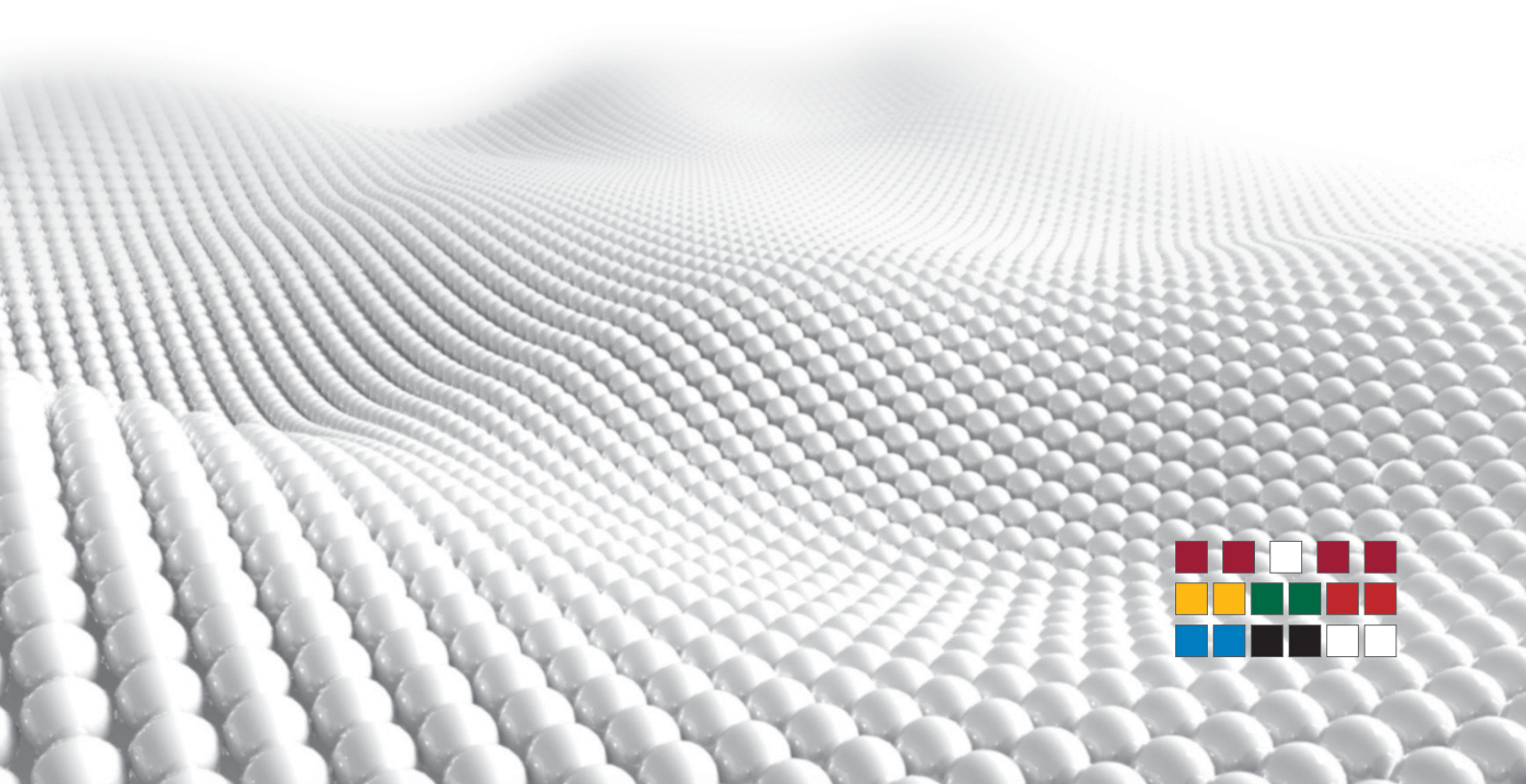
ISSN 2255-8950 (Online)
ISSN 2255-8942 (Print)

Volume 12(2024)

No. 3

Baltic Journal of Modern Computing



CO-PUBLISHERS

Vilnius University

UNIVERSITY OF LATVIA

Latvia University of Life Sciences and Technologies

Institute of Mathematics and Computer Science
University of Latvia

LIEPAJA UNIVERSITY

VIDZEMES AUGSTSKOLA

Baltic J. Modern Computing, Vol. 12 (2024), No.3, 222-247
https://doi.org/10.22364/bjmc.2024.12.3.01

On an Eigenplace Function – Mapping Relational to Absolute Space

Jurģis ŠĶILTERS, Līga ZARIŅA, Guntis Vilnis STRAZDS

University of Latvia, Laboratory for Perceptual and Cognitive Systems, Faculty of Computing, Raina blvd. 19, Rīga, Latvia

jurgis.skilters@lu.lv, liga.zarina@lu.lv,
guntis_vilnis.strazds@lu.lv

ORCID 0000-0002-3235-970X, ORCID 0000-0003-1799-3339, ORCID 0009-0001-0718-4397

Abstract. In this study we provide a descriptive framework of an Eigenplace function that maps from the relational space between objects to the numerical coordinate space that is a part of external reality. Our approach assumes that the Eigenplace function maps cognitively valid space (the relational structure linked with temporal intervals) to a mathematical set of coordinates. After a description of a detailed spatial ontology, some generalizations are discussed.

Keywords: space, time, region, vagueness, motion, Figure and Ground objects.

1. Introduction

What is the location of an object in space? Without examining the definition of object at this point, this is the question underlying the Eigenplace function in its simplest form. Important in this assumption is that, on the one hand, there are relations between objects that we mentally represent (normally by linking them to intervals in time or events), but, on the other hand, these objects are linked with concrete parts or regions of reality that are represented by sets of coordinates.

The Eigenplace function has been applied and examined in different fields ranging from geometry, mathematical models of spatial relations, and Qualitative Spatial Reasoning (Galton, 2000, Galton and Hood, 2005, Hood and Galton, 2006), to human geography (Golledge, 1992) and formal semantics of natural language (e.g., Kracht, 2002, Mador-Haim and Winter, 2015, Piñón, 1993, Pustejovsky, 2013, Wunderlich, 1991, Zwarts and Winter, 2000). However, there are no systematic frameworks converging different perspectives on this function. In the current study we aim at providing a descriptive and cognitively valid and inclusive framework, bringing together most of the current perspectives but also containing an ontology that can be implemented in an axiomatic way (in our case, by using a region-calculus), and corresponding to experimental evidence from research on spatial cognition.

In this paper, we first introduce the idea of the Eigenplace function and describe the constituents of a basic spatial ontology (e.g., regions, paths, and objects) and their main properties. Next, we describe the spatial relations between objects or regions according to the Region Connection Calculus (RCC) formalism extended with some additional operators. Further, the Eigenplace function is defined, and we describe how objects occupying regions in space and time and their relations can be mapped by using this function. This is followed by a description of uses of the Eigenplace function in modeling prepositions and Figure-Ground object roles, expressing them also in relation to the sequential order of time intervals and regions. The final sections of the paper provide remarks about time in Eigenplace functions, the modeling of vagueness and uncertainty in spatial reasoning, and the representation of motion and change of location. We conclude with general observations and a final discussion on the Eigenplace function.

2. Theoretical framework

A spatial configuration can be represented either relationally (which is the way space is cognitively represented) or absolutely (which is the way space is represented mathematically, and corresponds, e.g., to the Cartesian coordinate conception of absolute space). The idea behind the Eigenplace function is that each object in its relational sense is mapped to a concrete position in an absolute space.

An Eigenplace function as such does not explain where an object is, and does not even tell what the adjacent areas and objects are. Rather it presupposes the conception of an absolute space and maps relational information (crucial for spatial cognition) to the absolute (Cartesian or otherwise) space. Although every object has a unique location in absolute space, it is cognitively represented in relation to another object (reference object).

Canonically and informally the Eigenplace function can be defined as follows: every object stands in relation with other objects in a particular time interval and occupies a region in space (Wunderlich and Herweg, 1991, 758). This definition can be extended either by specifying the relations or adding other operators.

In our approach we assume (1) a particular ontology classifying types of constituents and (2) particular sets of relations. Both the ontology and relations – which are further explained below – can be flexibly extended.

2.1. Basic constituents

(cp. Mani and Pustejovsky, 2012, Talmy, 2000, Cohn et al., 1997)

A Basic Spatial Ontology consists of regions, paths, objects, orientation, distance and a few additional perceptual determinants (reference frame, manner and cause of movement). Below we describe these concepts and their main characteristics and principles in more detail. The constituents of the Basic Spatial Ontology are:

a) **regions (places)**;

b) **paths**; all segments of paths are also paths; all segments of paths are subsets or

elements of paths but not all paths are segments; also lines are considered as paths;

The basic principles that define paths and places and their mutual structure are as follows:

1. Paths are cognitively represented either as motions ('Mike walks') or as encoding a distinguished part (source, middle part or goal; e.g., 'Jim came from home', 'Black dog ran through the park', 'John went to a movie'). If a path contains a distinguished part, it is perceived *asymmetrically*.
2. *Spatial networks* are generated out of sets of paths (including their intersections) and regions adjacent to them.

c) objects

a. Figures (F; objects to be located);
b. Grounds (G; objects in virtue of which F are located);
c. Viewers (not always involved, and if involved not necessarily a part of the scene; however, location of a viewer (or an observer-induced axis) can determine the perception of Figure and Ground). For more on viewers see after the listing of the components of the ontology (see page 15).

There are several principles that characterize object properties and relations as perceived in places and paths:

1. Figures and Grounds are asymmetric in terms of (1) perceptual and functional prominence and dependencies, (2) their geometrical shape (Ground objects tend to be larger, more stationary; Figure objects – smaller, mobile), and (3) constraints of linguistic encodings (cp. Landau, 1996, Talmy, 2000, Carlson and Covell, 2005).
2. Borders, boundaries, and surfaces are considered as belonging to either regions or objects.
3. Physical objects are spatio-temporally persistent and their paths are connected: every physical object has a unique location and a continuous path in space and time without abrupt jumps, appearances and disappearances in different space or time segments (cp. Gardenfors, 2014, 128f.).
4. Objects are perceptually prior and primary with respect to regions; regions are perceived and discriminated as units of attention in virtue of their objecthood (Figure 1) (Scholl, 2001).
5. Objects are primary with respect to their locations, but the uniqueness of an object is possible in virtue of its location (cp. also Scholl, 2001, 14).

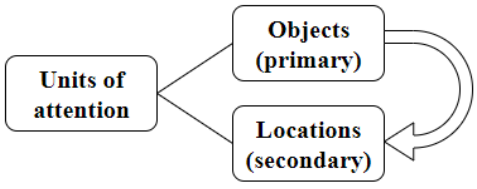


Figure 1. Units of attention (after Scholl, 2001, 14)

6. Events are also considered as objects (serving the role of either Figure or Ground), and only where it is necessary to distinguish between only spatial vs. only temporal readings, objects (in the narrower sense as a part of the spatial domain) can be contrasted with events (as a part of the temporal domain). In both cases, the relation between Figure and Ground holds. In general, the following regularities (Table 1) can be applied (cp. Wunderlich and Herweg, 1991, 760):

Table 1. Localizations and types of Figure (F) relative to a Ground (G) for spatial and temporal objects

Localization of F relative to a G		
Type of object	**F**	**G**
spatial	spatial object (or event)	spatial object
temporal	event (or spatial object)	event

This, however, means that the Eigenplace function would be different for places and events. Instead of an Eigenplace function

$$Eig: O \times T \rightarrow R$$

where O is a set of spatial objects, T – set of time intervals and R – set of spatial regions (places and paths), we would have another version for events

$$Eig_e: \mathcal{E} \times T \rightarrow R$$

where $\mathcal{E}$ is a set of events, T – set of time intervals and R a set of spatial regions (cp. Piñón, 1993). A crucial difference between Eigenplace with objects and Eingenplace with events is that we can never model the precise topology of events because each event has its own topology (if any at all) and eventually several topologies, whereas we can model precise topology of an Eigenplace with spatial objects. (For an alternative view concerning the application of topological relations to cognitive non-spatial relations cp. Lewin, 1936.)[1]

However, even when thinking of spatial objects, it is worth keeping in mind that spatial objects are transformed once they evolve in time (Jiang and Worboys, 2009).

7. In general there are different kinds of objects (cp. Gärdenfors, 2014, 129, Lyons, 1977, 442-445, Van Lambalgen and Hamm, 2005):
   (1) physical and spatial objects: they have a necessarily temporal embeddedness, and their locations are unique and their paths are continuous and connected trajectories; one and the same spatial object cannot be in two different places at the same time,
   (2) temporal objects (e.g., events): they have spatial constituents and they can also have sub-events; events can occur in several places at the same time (e.g., elections),

[1] Another tradition in considering events as objects is Davidsonian semantics (Pustejovsky, 2013, Davidson, 1969, for a different version cp. also Kim, 1973), assuming that the argument structure of a predicate contains a first-order individual *e*, i.e., $P(x_1,,x_n,e)$. Location of an event is a relation between the event variable *e*, and a location argument *l*, i.e., *loc(e,l).*

(3) abstract objects (e.g., propositions, sets): they do not have direct spatial or temporal properties.

d) orientation or direction (determining the relation between Figure and Ground): left, right, under, above.

e) distance: near/close, far; here we assume a relational conception of distance consisting of two main operators.

f) additional determining factors (neither inherently geometric nor topological):
   a. Frames of reference: not a part of a geometrically-topological framework but determining (d) above (the relation between Figure and Gound);
   b. Manner of movement;
   c. Cause of movement.

In a more general view, physical objects, events, shapes, and indeterminate regions are considered as first-order objects. Further, first-order objects can be classified according to different types – physical, temporal etc.

From a technical point of view, the present framework is largely consistent with Cohn et al., 1997, Randell et al., 1992, Kontchakov et al., 2010, Galton, 2014, in the way that their *sorts in the first-order sorted logic* correspond to the basic primitives in the sense of the present paper, i.e., 'regions' correspond to the sort REGION, 'objects' (Figures and Grounds) correspond to the sort PhysObj. Sort NULL referring to spatially non-existent objects is not represented in the current approach. The further axiomatization is based on the Region Connection Calculus 8 (RCC-8) (Randell et al., 1992) but enriched with several derived and non-derived relations.

2.2. Basic topological and geometric non-functional relations

Spatial objects (regions, paths, objects) are mutually situated in spatial relations that can be characterized topologically or geometrically. Below we describe the basic non-functional topological and geometric relations and their properties. Non-functional relations mean that the differences based on spatial prominence (distinction between Figure and Ground), frequent interaction, experience and general knowledge are not included in this ontology. The basic topological and geometric non-functional relations are:

1. ***Connectedness (C)*** between regions or objects which is the core relation underlying other spatial relations (cp. Cohn et al., 1997, Cohn et al., 1995; for topological interpretations: Galton, 2000, 82f.).[2]

C(x,y): *x* connects to *y*

[2] For the predecessor of this conception cp. Randell et al., 1992, Clarke, 1981, Clarke, 1985; for a discussion of the connection relation in context of temporal relations see: Galton 2009. A prominent axiomatic framework assuming connection as a foundational relation is that of B.L. Clarke (1981, 205) arguing that individual variables are spatio-temporal regions bound by two-place predicate 'connected with'. Informal idea about the foundational role of connectedness is also expressed by De Laguna (1922).

Connectedness is
(1) Reflexive: $\forall x[C(x,x)]$;
(2) Symmetric: $\forall x \forall y[C(x,y) \rightarrow C(y,x)]$.

Distance between objects bound by *C(x,y)* is zero.

If using a classical topological representation, we can define regions *x* and *y* as connected if their closures have at least one shared point:

$$C(x,y) \equiv_{def} cl(x) \cap cl(y) \neq \emptyset$$

In case of sets in $\mathbb{R}^n$ (sets in an n-dimensional vector space over real numbers) a set *S* is connected if between any two points of *S* there is a continuous path within *S* (Galton, 2000, 147).[3] In a wider sense, the primitive concept in our approach is a connection structure $(\mathcal{R}, C)$ where $\mathcal{R}$ is an arbitrary non-empty set of regions and C a symmetric binary relation on $\mathcal{R}$. The idea of connection structure is based on Whitehead's approach and further developed, made more precise by B.L. Clarke (Gerla, 1995) and enables to define inclusion '$\leq$' such that

$$x \leq y \Leftrightarrow C(x) \subseteq C(y)$$

Further, overlapping '$\mathbf{O}$' is such that

$$x\mathbf{O}y \Leftrightarrow \exists\, z \text{ such that } z \leq x \wedge z \leq y$$

Nontangential inclusion '$\ll$' would mean

$$x \ll y \Leftrightarrow C(x) \subseteq \mathbf{O}(y)$$

such that for every $z \in R$, $\mathbf{O}(z)$ is $C(z)$.

Apart from symmetry of connection relation (A1), the following axioms apply: there is no maximum for $\subseteq$ (A2); for every *x* and *y* there is a *z* that is connected to *x* and *y* (A3); connection is reflexive (A4); $C(x) = C(y) \Longrightarrow x = y$ (A5); any region *z* contains regions *x* and *y* that are not connected (A6) (Gerla, 1995, 1020, 1022).

According to RCC-8 (Figure 2) (Randell et al., 1992, Mani and Pustejovsky, 2012, 31) and enriched by some further relations (Cohn et al., 1997)[4] the following basic relations derived from *C* can be distinguished:

2. ***Disconectedness (DC)***: Regions or objects *A* and *B* do not touch each other; *A* is disconnected from *B*; *DC(A,B)* or formally defined substituting *A* and *B* by the variables *x* and *y*

$$DC(x,y) \equiv_{def} \neg C(x,y)$$

[3] Strictly speaking when two regions are connected within RCC: (a) they share (at least) a common point, or (b) their closures share a common point, or (c) distance between both regions is zero (Dong, 2008, 321, Cohn and Varzi, 2003). A point in RCC can be regarded either as a region or a special case or sort of a region; in the latter case it would be a categorically different object than the region. A more detailed discussion is an issue of another study (but cp. Dong, 2008) but a simple version of the mentioned definitions could be paraphrased by replacing point with region.

[4] For a context cp. also Bennett and Düntsch, 2007, Galton, 2004, Cohn and Renz, 2008, for an extension with Boolean operators cp. Wolter and Zakharyaschev, 2000, Stell, 2000; for a version containing distance and size relations cp. Dong, 2008. Topological and size information is integrated also in the approach by Gerevini and Renz, 2002. Another extension with direction relations is provided by Dube, 2017, Cohn, Li, Liu and Renz, 2014. For a relation-algebraic approach to RCC cp. Düntsch, Wang and McCloskey, 2001.

A topological interpretation of $DC(x,y)$ is $cl(x) \cap cl(y) = \emptyset$ where $cl(x)$ and $cl(y)$ are closed sets.

3. *Part (P)*: A region or object *A* is a part of a region or object *B*; *P(A,B)* or formally defined

$$P(x,y) \equiv_{def} \forall z[C(z,x) \rightarrow C(z,y)]$$

Parthood is
(1) Reflexive: $P(x,x)$,
(2) Transitive: $P(x,y) \wedge P(y,z) \rightarrow P(x,z)$. [5]

A topological interpretation of $P(x,y)$ is $x \subseteq y$.
An inverse version of *P* is also possible

$$Pi(x,y) \equiv_{def} P(y,x)$$

4. *Proper part (PP)*: A region or object *A* is a proper part of a region or object *B* whereby *B* unambiguously includes *A* as its part; *PP(A,B)* or formally defined

$$PP(x,y) \equiv_{def} P(x,y) \wedge \neg P(y,x)$$

A topological interpretation of $PP(x,y)$ is $x \subset y$.
An inverse version of *PP* is also possible

$$PPi(x,y) \equiv_{def} PP(y,x)$$

5. ***Overlap (O)***: A region or object *A* entirely overlaps with a region of object *B*: *O(A,B)* or formally defined

$$O(x,y) \equiv_{def} \exists z[P(z,x) \wedge P(z,y)]$$

A topological interpretation of $O(x,y)$ is $x \cap y \neq \emptyset$.

6. *External connectedness (EC)*: Regions or objects *A* and *B* touch each other at boundaries, i.e., are externally connected; *EC(A,B)* or formally defined

$$EC(x,y) \equiv_{def} C(x,y) \wedge \neg O(x,y)$$

Two regions or objects that touch each other are also called adjacent (Tomko and Winter, 2013, 181).

A topological interpretation of $EC(x,y)$ is $\partial x \cap \partial y \neq \emptyset \wedge x \cap y \neq \emptyset$, where ∂x and ∂y are borders of regions *x* and *y* respectively.

Alternatively if we indicate bounded regions, i.e., their interiors (x^o, y^o), we can define *EC* as $x \cap y \neq \emptyset \wedge x^o \cap y^o = \emptyset$ (cp. also Li and Cohn, 2012).

7. *Partial overlap (PO)*: Regions or objects *A* and *B* partially overlap each other in space; *PO(A,B)* or formally defined

$$PO(x,y) \equiv_{def} O(x,y) \wedge \neg P(x,y) \wedge \neg P(y,x)$$

[5] According to the default interpretation we assume that these formulae are universally quantified; we are omitting universal quantifiers here and elsewhere for the sake of simplicity (cp. also Galton, 2014, 293.)

A topological interpretation of $PO(x,y)$ is $x \cap y \neq \emptyset \wedge x \nsubseteq y \wedge y \nsubseteq x$.

If bounded regions (their interiors are indicated) then *PO* is $x^o \cap y^o \neq \emptyset \wedge x \nsubseteq y \wedge x \not\supseteq y$ (cp. also Li and Cohn, 2012).

8. ***Equality (EQ)***: Regions or objects *A* and *B* occupy the same space (they are spatially identical); *EQ(A,B)* or formally defined

$$EQ(x,y) \equiv_{def} P(x,y) \wedge P(y,x)$$

A topological interpretation of $EQ(x,y)$ is $x = y$.

9. ***Discreteness (DR)***: Regions or objects *A* and *B* are discrete from each other; *DR(A,B)* or formally defined

$$DR(x,y) \equiv_{def} \neg O(x,y)$$

Discreteness can also be expressed as either disconnectedness or external connectedness, i.e.,

$$DR(x,y) \equiv_{def} EC(x,y) \vee DC(x,y)$$

10. ***Tangential proper part (TPP)***: Region or object *A* is inside the region or object *B* and *A* touches the boundary of *B*; *TPP(A,B)* or formally defined

$$TPP(x,y) \equiv_{def} PP(x,y) \wedge \exists z[EC(z,x) \wedge EC(z,y)]$$

A topological interpretation of $TPP(x,y)$ is $x \subset y \wedge \partial x \cap \partial y \neq \emptyset$.

If boundedness of regions and their interior parts are taken into account we can write $x \subset y \wedge x \not\subset y^o$ where y^o stands for bounded region (cp. also Li and Cohn, 2012).

11. ***Non-tangential proper part (NTPP)***: Region or object *A* is inside the region or object *B* and does not touch the boundary of *B*; *NTPP(A,B)* or formally defined

$$NTPP(x,y) \equiv_{def} PP(x,y) \wedge \neg\exists z[EC(z,x) \wedge EC(z,y)]$$

A topological interpretation of $NTPP(x,y)$ is $x \subset y \wedge \partial x \cap \partial y = \emptyset$. And interpretation where bounded / interior parts are taken into account $x \subset y^o$(cp. also Li and Cohn, 2012).

From *TPP* and *NTPP* directly derived relations that will not be further explored in this paper:

(1) ***Tangential proper part inverse (TPPi)***: Region or object *B* is inside the region or object *A* and *B* touches the boundary of *A*.

(2) **Non-*tangential proper part inverse (NTPPi)***: Region or object *B* is inside the region or object *A* and *B* does not touch the boundary of *A*.

The Basic principles holding for relations 1-11 relate to symmetry. Most of the relations – *C*, *DC*, *DR*, *O*, *PO*, *EC*, *EQ* – are symmetric. The relations *P*, *PP*, *TPP*, and *NTPP* are not symmetric and can have an inverse interpretation (cp. Galton, 2009, 179).

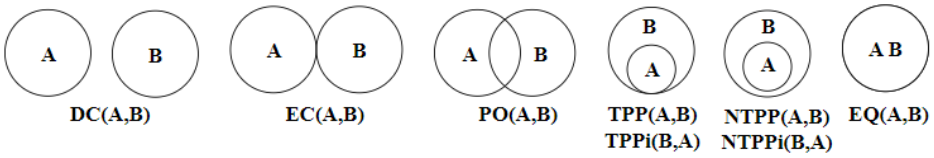


Figure 2. Illustration of RCC-8 relations

The so far described relations are basic topological relations respective to RCC-8. However they can be enriched. A useful relation for expressing everyday contexts is ***convex hull (conv)*** (Cohn et al., 1997, 287ff.; Cohn, 1995; Cohn et al., 1995, a critical discussion cp. Dong, 2008, 349f.).

Convex hull of *x* is a function ***conv(x)*** that can be considered as a spatial primitive[6] referring to the smallest convex region that includes *x*. "A convex region can be defined as one having such a shape that a straight line joining any two points within the region does not go outside it. The *convex hull* of an arbitrary region is then the smallest convex region that contains it[.]" (Cohn et al., 1998, 8)

$$conv(x) \equiv_{def} EQ(x, conv(x))$$

which, in turn, means, e.g., that

$$TPP(x, conv(x));$$
$$P(x,y) \rightarrow P(conv(x), conv(y)).$$

The elementary properties of *conv* (cp. Galton, 2000, 182) are as follows*:*

$$x \subseteq conv(x);$$
$$x \subseteq y \rightarrow conv(x) \subseteq conv(y);$$
$$conv(x \cap y) \subseteq conv(x) \cap conv(y);$$
$$conv(x) \cup conv(y) \subseteq conv(x \cup y).$$

Conv(x) enables to define regions that are entirely/partly inside or outside the convex hull of *x* but not overlapping *x* (Figure 3) (Cohn et al., 1997, 288, Randell et al., 1992):

(1) inside (*inside*)

$$inside(x,y) \equiv_{def} DR(x,y) \wedge P(x, conv(y))$$

or

$inside(x,y) \equiv_{def} \neg P(x,y) \wedge P(x, conv(y))$ (Cohn et al., 1995, 836)

(2) **partly inside (*p_inside*)**

$$p\_inside(x,y) \equiv_{def} DR(x,y) \wedge PO(x, conv(y))$$

or

$p\_inside(x,y) \equiv_{def} \neg P(x,y) \wedge PO(x, conv(y)) \wedge \exists w[P(w, conv(y)) \wedge \neg P(w,y) \wedge PO(w,x)]$ (Cohn et al., 1995, 836)

(3) **outside (*outside*)**

$$outside(x,y) \equiv_{def} DR(x, conv(y))$$

[6] Thus, the underlying formal theory contains two primitive relations *C*(*x*,*y*) and *conv*(*x*).

or
$outside(x,y) \equiv_{def} \neg P(x,y) \wedge \neg\exists w[P(w,conv(y)) \wedge \neg P(w,y) \wedge PO(w,x)]$ (Cohn et al., 1995, 836).

Also inverse relation of convexity can be formulated.

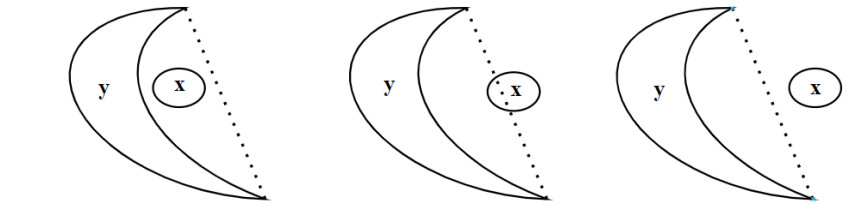


Figure 3. Regions that are a) entirely inside, b) partly inside and c) outside the convex hull

In general, two interpretations of the relation ***inside(x,y)*** should be distinguished: a topological and geometrical (Randell et al., 1992): in the former case (***top_inside***) a region or an object is inside of another region or object if it is a proper part with an surrounding region or object, i.e., there is no cut through the surrounding region or body. In geometrical case (***geo_inside***) an object or a region is inside another but excluding the topological containment.

Accordingly:

$$top\_inside(x,y) \equiv_{def} inside(x,y) \wedge \forall z\big[[conv(z) \wedge C(z,x) \wedge C(z,outside(y)] \rightarrow O(z,y)\big];$$

$$geo\_inside(x,y) \equiv_{def} inside(x,y) \wedge \neg top\_inside(x,y).$$

Thus, topological insideness (containment) has to be distinguished from geometrical insideness that contains convex hull relations as its subsets. In the latter case, relations referring to inside, partial inside and outside have to be distinguished.

Keeping in mind the differences between geometric and topological containment, at least three geometrically and topologically distinct types can be distinguished (Figure 4) (Zwarts, 2017, 14):

(1) **topological enclosure** where regions are related by *TPP* or *NTPP* (and accordingly the inverse relations). E.g., 'Honey in a closed jar', 'A bug in an amber';
(2) **convex geometrical enclosures** where partial geometric enclosure ($p\_inside(x,y)$) is the most prominent instance. E.g., 'A flower in a vase';
(3) **scattered geometric enclosure**: enclosure is perceived without any topological connectedness or containment. E.g., 'Birds in the trees'.

The convexity relation enables to express configurations of inclusion (alternatively they can also be expressed in Wunderlich's (1993, 124ff.) framework using the so-called focusing effects), where there is a partial inclusion (represented by 'in') of F in G and the verb expresses a supportive or holding function of G, whereby only a part of F is in functional interaction with G and is thus highlighted. But the whole spatial configuration

in these cases can be plausibly expressed using convexity relation. Convexity regions frequently contain regions of functional interaction. E.g., 'The pipe (F) held in the mouth (G)', 'The stick (F) held in the hand (G)'.

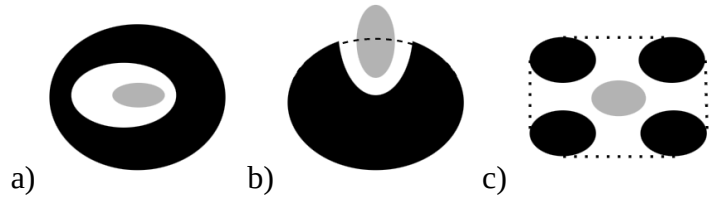


Figure 4. Three types of 'in' in RCC: a) total topological enclosure, b) partial geometric enclosure, and c) scattered geometric enclosure (after Zwarts, 2017, 14)

Finally, a relation that can be expressed in applying and combining (1)-(11) relations is **betweenness (*Betw*)**. Object or region *B* is between *A* and *C* if it is also between *C* and *A* (Miller and Johnson-Laird, 1976, 61), i.e., $Betw(A,B,C) \leftrightarrow Betw(C,B,A)$.

3. Eigenplace function

As argued before, Eigenplace is a universal and default relation that holds in all spatial relations covering static, locational and dynamic, directional configurations, meaning that all objects (together with temporal segments) are mapped to spatial regions. If the temporal dimension is added then we might define: Every object in time is located in a concrete place (in relation with other objects and within the ontology that we are proposing) and there are no two objects occupying the same place at the same time. If *O* is a set of arbitrary objects, *T* is a set of time intervals and *R* is a set of regions (exact locations in, e.g., a coordinate system), then the simplest versions of the Eigenplace function are:

$$Eig: O \rightarrow R$$
$$Eig^t: O \times T \rightarrow R$$

Accordingly, an atemporal and a temporal version of Eigenplace have to be distinguished.

The idea of Eigenplace corresponds also to the fundamental principle in geography that there are no two discrete things that occupy the same region in space at the same time (cp. Golledge, 1992, 205).

In what follows we will, first, explore a more restrictive analysis of spatial prepositional relations and their Eigenplace mappings and then move on to a more general approach to Eigenplace relations.

In combining relative and absolute representations we can write

$$Eig: O \rightarrow R_{CART}$$

meaning that a set of objects (*O*) are always located in concrete place (R_{CART}) on a Cartesian plane (although other kinds of coordinate representations are possible).

If we assume that spatial objects are linked to concrete temporal units (*T*), either points or intervals,

$$Eig: O \times T \rightarrow R_{CART}$$

further we assume that

$$r_1, \dots, r_n \in R_{CART}$$
$$o_1, .., o_n \in O: O \neq \emptyset$$
$$t_1, .., t_n \in T$$

if *Rel* is a subset of spatial relations (RCC extended with some geometric and functional primitives) then

$$Eig: \langle Rel, o_1, \dots, o_n \rangle \times t_i \rightarrow r_j$$

What are canonical options for the ways in which objects can occupy regions in space? According to Galton (2000, 168-170), if *o* is an object and *r* – a region, there are several sets of possible relations:

1. *o* and *r* are congruent (*r* is a possible value of *o*): *DC, EC, PO, EQ*
2. *o* just fits into *r: DC, EC, PO, TPP*
3. *o* covers *r*: *DC, EC, PO, TPPi*
4. *o* can fit right inside *r*: *DC, EC, PO, TPP, NTPP*
5. *o* covers more than *r: DC, EC, PO, TPPi, NTPPi*
6. *o* and *r* are incommensurate (none of above relations holds): *DC, EC, PO*

We might express our framework in terms of Galton (2000) by writing $Rel(Eig(a), Eig(b))$ where *Rel* is a spatial relation and *Eig* are Eigenplace of Figure (a) or Ground (b). E.g., $DC(Eig(a), Eig(b))$, to denote an object *a* that is outside of another object *b;* further *DC* – disconnectedness relation and *Eig* – the Eigenplace function mapping each object to a concrete place.

In general

$$Rel(Eig(a), Eig(b))$$

where Rel is one of the canonical *RCC* relations (except *EQ*). This means that an object *a* stands in a certain relation to another object *b* if the position of *a* stands in this relation to the position of object *b* (Galton, 2000, 168). In total, it means

$$Rel(Eig(o_1), \dots, Eig(o_n)) \Leftrightarrow Rel(o_1, \dots, o_n) \rightarrow r_j$$

As for the relation *EQ*, we might write

$$EQ(o, r): o \times t \rightarrow r$$

where *o* is an arbitrary object, *t* – a time and *r* – a region (place), which is in turn equivalent to the canonical Eigenplace relation.

To sum up so far, the general schema of the Eigenplace function is

$$\mathcal{R}(o_1, \dots, o_{n:n>1}) \times t_i \rightarrow r_j$$

where $o_1, \dots, o_n$ are arbitrary objects and t_i time intervals, and r_j the corresponding region in real (e.g., Cartesian) space; $\mathcal{R}$ is any arbitrary spatio-temporal relation. To put it more precisely – spatial relations between objects in time occupy a particular region in

coordinate space

$$\langle \mathcal{R}, o_1, \dots, o_{n:n>1} \rangle \times T \rightarrow R$$

Different spatial relations R_1^s and R_2^s between the same set of objects occur either in the same or different temporal intervals and different locations in coordinate space

$$\langle R_1^s, o_1, \dots, o_{n:n>1} \rangle \times t_i \rightarrow r_1$$
$$\langle R_2^s, o_1, \dots, o_{n:n>1} \rangle \times t_j \rightarrow r_2$$

It should be kept in mind that spatial relations $\langle R_1^s, \dots, R_n^s \rangle$ are ordered with respect to canonical RCC relations and generalized relations in such a way that $Rel \subset R^S \subseteq \mathcal{R}$.

Another subset of relations is temporal relations R^t such that $R^t \subseteq \mathcal{R}$. We assume that R^toperate on the set of temporal intervals and we also assume that R^t are Allen interval relations (Allen, 1983) such that a temporal structure is $\langle R^t, t_1, \dots, t_{j:j>1} \rangle$.

4. Classical approaches: Eigenplace in prepositions and Figure-Ground roles

In a narrower sense, an Eigenplace function is an intuitively plausible and formally clear approach to the analysis of relations between Figure and Ground (Wunderlich, 1991, 597, Asbury et al., 2008, 12, Zwarts, 1997, Svenonius, 2010, Mador-Haim and Winter, 2015) yielding for every "object or event the place it occupies (its 'Eigenplace'), which is some region" (Wunderlich, 1991, 597). Eigenplace functions are related so that every object is mapped to a particular place in space and all objects are directly or indirectly related in space.

Classically D. Wunderlich provides a *general scheme* holding for all spatial configurations and *particular schemes* describing some specific configurations. A general scheme for prepositional information says that

$$\langle F, G \rangle \in [\![Prep]\!] \text{ iff } Eig[F] \subseteq R_{Prep}[G]$$

where *F* and *G* are Figure and Ground, *Prep* is a spatial preposition or other spatial expression, R_{Prep} is a neighbourhood function for the preposition *Prep*, and Eig is the Eigenplace function.

The important assumption by Wunderlich is that *F*-objects and *G*-objects are paired only indirectly in using the neighbourhood region of *G* (Wunderlich, 1991, 598). The neighborhood region of *G* (*search domain*) is sensitive to perceptual, geometrical, reference frame and functional properties, whereas the properties of *F* determine whether *F* can be included in the neighborhood region of *G*, i.e., $R[G]$.

If a function *INT*[*G*] yields a set of regions *internal* to the Ground, then the 'in' location is

$$\langle F, G \rangle \in [\![in]\!] \text{ iff } Eig[F] \subseteq INT[G]$$

The Eigenplace function complemented with some other basic functions – such as *EXT*[*G*] (referring to regions *external* to *G*), *PROX*[*G*] (referring to regions in the *proximity* of *G*) and additional markers such as *axis-orientation* $\pm VERT$ – allow one to represent other spatial relations as well. E.g.,

$$\langle F, G \rangle \in [\![under]\!] \text{ iff } Eig[F] \subseteq EXT[G, -VERT].$$

Wunderlich also introduces a predicate *LOC(x,r)*, meaning that an object *x* is located in a region *r* and, thus, in fact, fulfills the role of Eigenplace function; and in a more general schema $LOC(F, r[G])$, with the meaning that *F* is located in the region *r* with respect to *G* (Wunderlich, 1991, 598, Wunderlich, 1993, 113, cp. also Zwarts, 1997, 60f. for a different framework). The basic truth condition here is:

$$LOC(F, r) \text{ is true iff } L[x] \subseteq r$$

where *L* is the Eigenplace of *x* (place occupied by *x*) and '⊆' spatial containment (Wunderlich, 1993, 114).

Practically, this makes it possible to express Eigenplace-relations within a lambda-formalism:

$$\langle F, G\rangle \in [\![under]\!] \text{ iff } Eig[F] \subseteq EXT[G, -VERT] \text{ is identical to}$$
$$\lambda G \lambda F\ LOC(F, EXT[G, -VERT]) \text{ and}$$
$$\langle F, G\rangle \in [\![in]\!] \text{ iff } Eig[F] \subseteq INT[G] \text{ is identical to}$$
$$\lambda G \lambda F\ LOC(F, INT[G])$$

Accordingly, the general scheme is

$$\lambda G \lambda F\ (LOC(F, r[G]) \wedge \mathcal{C}(F, G))$$

where $\mathcal{C}$ refers to additional constraints (e.g., relations of contact, intersection, enclosure) (Wunderlich, 1991, 599, Wunderlich, 1993, 114).

Within this framework the formal representation of basic locational prepositions can be described as follows (Wunderlich, 1993, 113):

'in' $\lambda G \lambda F\ LOC(F, INT[G])$
'by' $\lambda G \lambda F\ LOC(F, EXT[G])$
'over' $\lambda G \lambda F\ LOC(F, EXT[G, +VERT])$
'under' $\lambda G \lambda F\ LOC(F, EXT[G, -VERT])$
'in front of' $\lambda G \lambda F\ LOC(F, EXT[G, +obs])$
'behind' $\lambda G \lambda F\ LOC(F, EXT[G, -obs])$

where '*obs*' means dependence on observer axis: '$+obs$' directed toward the observer and '$-obs$' directed away from the observer.

In some path-expressions Wunderlich adds a dimension parameter *D[F]* that is relative to the movement of the figure on a path and also additional relations – *ENCL* (to enclose *G*), *INTERSEC* (to intersect *G*), to be parallel to the maximal axis of *G* (*PARAL*; *MAX*). The EXT and *INT* are defined as $EXT = proximal\_exterior\_of$ and $INT = interior\_of$, and $PROX = EXT \cup INT$ (Wunderlich, 1993, 115-118). These relations allow to model such basic locational prepositions as:

'around' $\lambda G \lambda F\ LOC((F, EXT[G]) \wedge ENCL(D[F], G))$
'through' $\lambda G \lambda F\ LOC((F, INT[G]) \wedge INTERSEC(D[F], G))$
'along' $\lambda G \lambda F\ LOC((F, PROX[G]) \wedge PARAL(D[F], MAX[G]))$

A further important property is that all *objects are located relative to other objects*, i.e., every object has a *neighborhood* of other objects. If *O* is a set of objects, *T* – set of time intervals and *R* – set of regions then there is a family U_j of neighborhood-functions:

$$U_j = \{u_j : O \times T \to R\}, j \in N$$

where $u_j(a,t)$ is a special (concrete) neighborhood of an object *a* at time *t.* Object *b* can be localized relative to *a* according to a neighborhood-function u_j such that

$$p(b,t) \sqsubseteq u_j(a,t)$$

where $\sqsubseteq$ is a spatial part-of-relation. According to $p(b,t) \sqsubseteq u_j(a,t)$ we can say that the place of object *b* is a part of the j-neighborhood of the object *a* (cp. Wunderlich and Herweg, 1991, 759, 760). Usually the object *b* serves the role of focal object (Figure), whereas *a* is the reference object (Ground) that enables the localization of *b*, i.e., $p(F,t) \sqsubseteq u_j(G,t)$.

In a more abstract way according to Wunderlich and Herweg (1991, 772f.) we can introduce a general localization relation *LOC* and, thus, the relations between every two spatial objects can be expressed as either

$$\lambda x \lambda y\ LOC(x, u_j(y)) \text{ or}$$
$$\lambda x \lambda y\ LOC[p(x) \sqsubseteq u_j(y)]$$

where $LOC(x,R)$ is a general localization relation with the meaning that the place of an individual *x* is a spatial part of a spatial region *R*; U_j is a family of functions u_j assigning certain neighborhoods to individuals and *p* is a localization function assigning places to individuals. According to Wunderlich and Herweg, the neighborhood relation U_j contains specific differences between spatial relations between objects (in our case, the specific differences are expressed using basic topological and geometric non-functional relations together with orientation and distance primitives. E.g., the meaning of a spatial preposition 'on' can be expressed

$$ON(x,y) \leftrightarrow LOC(x, ON^*(y))$$

where ON^* is a specific neighborhood function characterizing 'on'. More generally

$$\lambda y, \lambda x, LOC(x, ON^*(y))$$

or, using the Figure and Ground distinction,

$$\lambda y, \lambda x, LOC(F, ON^*(G))$$

Thus, the meaning of a spatial preposition is a localization relation between objects (in the case of the current approach, Figure and Ground objects). According to Wunderlich and Herweg (1991, 777), the core pattern of all locative prepositions is the following scheme:

$$\lambda y, \lambda x, LOC(x, PREP^*(y))$$

where *x* is the Figure and *y* the Ground, and $PREP^*$ is a characteristic neighborhood function of a *y* that distinguishes a preposition. E.g.,

$$\lambda y, \lambda x, LOC(x, IN^*(y))$$

is a general and schematic formal representation of the meaning of the preposition 'in'. The background intuition of $LOC(x, IN^*(y))$ is that there is a region $IN^*(y)$ that enables the localization of *x.* I.e., not just the Ground but also a special configurational part (characterized by *IN**) of it enables one to locate the Figure. The localization of the Figure (mapped to a spatial region) is enabled only by localization of a Ground (that is also mapped to a spatial region and in this case specified by a particular preposition,

$PREP^*(y)$) (Wunderlich and Herweg, 1991, 777).

A slightly different representation involving observer-induced axis (*d*) is the case of dimensional prepositions (e.g., describing the area 'in front of') (Wunderlich and Herweg, 1991, 778f.): e.g.,

$$BEVOR(x, y, d) \leftrightarrow LOC(x, BEVOR^*(y, d))$$

Thus, in general, objects – in accordance with the Eigenplace function – are always mapped on spatial regions and are related to each other. In certain cases additional constraints (e.g., based on functional knowledge) have to be applied $\mathcal{C}(x, y)$ (Wunderlich and Herweg, 1991, 777).[7] Such relations can be expressed in the framework of an Eigenplace relation

$$o \times \tau \rightarrow \mathcal{r}$$

where o is a type of object, τ – a type of time intervals and $\mathcal{r}$ – a type of spatially extended concrete regions (concrete regions in space). $b_1, \dots, b_n, e_1, \dots, e_n, r_1, \dots, r_n \in o$, where $b_1, \dots, b_n$ is a set of physical objects (e.g., cups, tables, houses), $e_1, \dots, e_n$is a set of events (e.g., birthday celebration, meeting) and $r_1, \dots, r_n$ is the set of regions that are not linked to a concrete spatial area (locationally indeterminate shapes, contours). E.g., 'Celebration party (event) was in the residential area (locationally indeterminate region) in front of the city council building (physical object)'. Paths are also a subset of region types $P_1, \dots, P_n \in \mathcal{r}$ and $t_1, \dots, t_1 \in \tau$.

A more robust formulation (involving also time intervals) of an Eigenplace function (called *Lokalisierungsfunktion p*) is provided by Wunderlich and Herweg (1991, 758): every object in a time interval occupies a region in space:

$$p: O \times T \rightarrow R$$

where *O* is a set of objects (whereby also events can be considered as objects; the difference is, however, that in the case of events there are no clear topological relations like in the case of physical objects), *T* – a set of time intervals and *R* – a set of regions and *p* is a certain place. This means that every place (i.e., *p(o,t)*) is a region that is occupied by an object *o* at time *t*.

This formulation of the Eigenplace function corresponds to the *loc'* function by Kracht (2008, 40, 2002, 179, cp. also Piñón, 1993):

$$loc': e \times \tau \rightarrow r$$

where *e* denotes a type of object, τ - type of time-points and r – type of regions. Function *loc'* generates a product of an object and a time point, and returns the region the object occupies at this time.

In expressions of directional spatial relations Eigenplace function refers to the sequential order of times and regions. According to Wunderlich and Herweg (1991), there is a path-function (*Wegfunktion w):*

$$w: O \times SeqT \rightarrow SeqR$$

where *O* is a set of objects, *T* – a set of time intervals, *R* – a set of regions, and *Seq* – a sequence relation of time intervals or regions; a, t_i is a region occupied at a certain time

[7] To distinguish from the relation connect, a slightly different symbol is used; initially Wunderlich and Herweg (1991,777) use C(x,y).

t_i, where $0 \leq i \leq 1$. Accordingly a, t_0 is the region at the beginning of a path and a, t_1 is a region occupied by an object at the end of a path (cp. Wunderlich and Herweg, 1991, 759, for an analysis of Eigenplace functions in vector space semantics cp. Zwarts and Winter, 2000, 175ff.). A general representation of paths in Eigenplace terms (corresponding to the *Wegfunktion w* by Wunderlich and Herrweg, 1991) is:

$$Path\_function: o \times Seq\ \tau \rightarrow Seq\ \mathcal{r}$$

To formalize this idea in terms consistent in the current approach: If *A* and *B* are objects or regions, *REL* is a subset of an extended version of RCC (including additional geometric features that are described before), and τ – a type of time intervals and $\mathcal{r}$ – a type of spatially extended concrete regions (concrete regions in space), then the spatially extended path referred to by *A* and *B* at a certain time is

$$REL(A, B) \times Seq\tau \rightarrow Seq\mathcal{r}$$

such that $\langle t_1, \dots, t_n \rangle \in \tau$, $(P_1, \dots, P_n) \in Seq\mathcal{r}$.

5. Remarks on time in Eigenplace

Let us assume time as consisting of intervals.[8] Intervals are linearly ordered and their relations can be constrained as discrete, dense, continuous, bounded or unbounded in each direction (Bennett and Galton, 2004, 16). A general temporal ordering is a History structure

$$\mathcal{H} = \langle S, T, \prec, H \rangle$$

where *S* is a set of states in the world: $s_1, \dots, s_n$. Further we assume that *S* is a relational structure consisting of at least extended RCC. *T* is a set of time intervals (or points): $t_1, \dots, t_n$; $\prec$ is irreflexive linear order on *T* (dense, discrete or continuous). *H* is a set of histories $h_1, \dots, h_n$, i.e., functions from *T* to *S:*

$$H: T \rightarrow S$$

such that $h_1: t_1 \rightarrow s_1$,…, $h_n: t_n \rightarrow s_n$.

We assume some additional functions to describe terminal parts of intervals: $beg(t_i)$ and $end(t_i)$ are functions referring to the beginning and end of an interval t_i.

Further, we agree with Bennett and Galton (2004) and assume a truth functional meaning: $[\![\alpha]\!]^{\mathcal{A}}_{h,t}$, i.e., denotation of expression α at an index $\langle h, t \rangle$ and according to the assignment $\mathcal{A}$ determining the values of non-logical constants: e.g. a set of all assignments for which an expression φ is true, i.e., a truth set *TS* (Bennett and Galton, 2004, 27f.):

$$[\![\varphi]\!]_{TS} = \left\{ \langle \mathcal{A}, h, t \rangle | [\![\varphi]\!]^{\mathcal{A}}_{h,t} = t \right\}$$

Events consist of intervals and relate to event types. One and the same event type can refer to several events. Intervals $\delta_1, \dots, \delta_n$ satisfy the event sequence $e_1, \dots, e_n$ but then $\delta_1, \dots, \delta_n$ has to satisfy the sequence of event types $e_1^*, \dots, e_n^*$ such that $e_i^* \subseteq e_i$

[8] We assume that points in both spatial and temporal senses are rather abstractions and special cases than actual parts of the perceivable world, therefore preferring non-atomistic intervals and regions and the basic constituents.

(Bennett and Galton, 2004, 42).

Next we would like to describe the Eigenplace of an event (cp. Pustejovsky, 2013): If an event is a structured object $\mathcal{E}$ where a relation *R* applies at time *t*, we can write $\langle R, o_1, \dots, o_n, t\rangle$, then the localization of an object in an event is $loc(o,t) = r_o$. An event with its object localizations is $\langle R, o_1, \dots, o_n, r_{o1}, \dots, r_{on}, t\rangle$, where $r_{o1}, \dots, r_{on}$ are object locations in space.

Normally spatial objects are transformed in time (there is even an approach assuming that events specified via the changes in topological structure are called topological events (Jiang and Worboys, 2009, 34)).

6. Representing vagueness and uncertainty in spatial reasoning

Sometimes we lack the necessary information to determine a spatial location and sometimes spatial objects are inherently vague (e.g., hills, swamps). When dealing with spatial uncertainty or vagueness, it has to be kept in mind that although every spatial object (also vague) has some precise extension in real world (although we do not know it or cannot adequately represent it), we still can use relational information to narrow down the area where the object can be located. We can frequently relationally specify an area where some objects are to be located. This area is a region, or relational structure referring to a concrete extension in the real world.

One way for dealing with vague and uncertain spatial information consistently with RCC-based formalisms is to use an ***anchoring relation***[9] as defined by Galton and Hood (Galton and Hood, 2005, Hood and Galton, 2006, Hood, 2007, for recent applications see: Vasardani et al, 2017, Chen et al., 2017, for an approach in formalization of common sense reasoning of containment in case of incomplete information: Davis et al., 2017; alternative approaches on approximate reasoning in RCC5 and RCC8: Bittner and Stell, 2000).

Anchoring relations enable one to define areas based on *what is known* instead of *specifying a precise location* (which is frequently impossible because of a lack of information). Further, there are at least two ways in which spatial information can be indeterminate: (a) the spatial object we are dealing with might be *vague* (i.e., we cannot define a precise border for it; e.g., hills, forests are instances of spatial objects where they might gradually cease to exist or transform into other spatial objects), (b) spatial information can be *uncertain*, i.e., we might not have enough knowledge to describe the object (see Hood and Galton, 2006, Hood, 2007).

According to this approach we can refer to a known area that in turn includes a region that is indefinite within this area. E.g., we know that an accident occurred in an area where two districts intersect (and if *Distr* stands for a district and *t* for time interval and *r* for a concrete region in real world) then:

$$PO(Distr1, Distr2) \times t_i \rightarrow r_a$$

However, the exact location of the accident $Accident(r_b)$ within r_a is not known. If $r_b \subset r_a$ and if it is known that it occurred somewhere in front of two houses we can write:

$$IN\_FRONT\_OF(Accident, DC(House1, House2)) \times t_i \rightarrow r_b$$

[9] The anchoring relation basically corresponds to the Eigenplace relation.

However, we do not know the exact coordinates of r_b. (This is the reason why we write r_b^V or r_a^V to indicate that r_a or r_b is vague in epistemic terms. I.e.,

$$IN\_FRONT\_OF(Accident, DC(House1, House2)) \times t_i \rightarrow r_b^V)$$

The idea behind the anchoring approach is that there are two different spatial structures:

a. *information space* – information regarding spatial objects, locations and their relations to each other. Information space is expressed in a relational language (i.e., language that is sufficiently rich to allow expressing spatial relations). Information space also contains non-spatial information (e.g., temporal, emotional, social).
b. *precise space* – consisting of exact locations of objects as expressed in a numerical coordinate system (e.g., Cartesian system). Exact space corresponds to an extensional point set topology in a coordinate system.

Information space and precise space are related by mapping information space to precise space in a way that allows *more than one type of relation in information space* to correspond to *one and only one region in precise space*. There are several different ways in which objects in information space can be linked to precise space.

If *R* is a spatial relation (e.g., one of the extended RCC relations) applying to a set of objects $o_1, \dots, o_n$, $\mathfrak{C}$ – precise space (e.g., Cartesian coordinate space) and r_k – a region of it (such that $r_k \in \mathfrak{C}$) then

$$R(o_1, \dots, o_n) \times t_i \rightarrow r_k$$

Possibilities of *anchoring* according to Galton and Hood, 2005; Hood and Galton, 2006; if $o_1, \dots, o_n \in O$ and $r_k \in \mathfrak{C}$ are:
An object o_i is anchored <u>in</u> r_k means that o_i is located within/inside r_k;
An object o_i is anchored <u>over</u> r_k means that r_k falls within the location of o_i (the location of object o_i contains the whole r_k);
An object o_i is anchored <u>outside</u> r_k means that there is no part of o_i that is located inside of r_k;
An object o_i is anchored <u>alongside</u> r_k means that o_i abuts r_k.

These anchoring relations $\mathbb{A}(O, \mathfrak{C})$ are relating sets of objects (O) in relational space with sets of locations (exact regions) in precise space $\mathfrak{C}$. The intuition behind this is that objects are always located in precise regions even if we do not know exact location.

The idea of anchoring is plausible since we usually talk about objects relationally and use vague and uncertain concepts even though objects do have exact locations (even if we do not know them, which is usually the case). Assuming $r_j, r_k \in \mathfrak{C}$ and $o_i, o_j \in O$, and *loc* is a function denoting the location of an object ($loc \in \mathbb{A}$), we can say according to Hood and Galton (2006) that two constraints apply to anchoring:

(1) if an object is anchored over a region then this region is a part of any region this object is anchored in

$$(in, r_j) \in loc(o_i) \wedge (over, r_k) \in loc(o_i) \rightarrow r_k \subseteq r_j \text{ and}$$

(2) there are no two regions in which an object is anchored such that they are disjoint

$$(in, r_j) \in loc(o_i) \wedge (in, r_k) \in loc(o_i) \rightarrow r_k \cap r_j \neq \emptyset$$

An approach where anchoring is applied to the analysis of preposition 'at' is provided by Vasardani et al., (2017). Imagine the utterance: 'Let us meet at the park'. The meeting point is anchored either (a) inside, (b) along its boundaries, or (c) close to but outside the park. According to Vasardani et al. (2017), a Figure object $F \in o_1, \dots, o_n$ is at Ground object (anchoring area) if and only if *F* is anchored in the region r_j by Ground object $G \in o_1, \dots, o_n$.

$$R(o_1, \dots, o_n) \times t_i \rightarrow r_j$$

Accordingly:
F is in *G* if *F* is anchored in the region by *G*;
F is near *G* if *F* is anchored in the relative complement of *G*.

If $\mathbb{r}$ is anchoring relation $\mathbb{r} \in \{in, over, alongside, outside\}: \mathbb{r} \subseteq \mathbb{A}$ then anchoring happens as an ordered pair $\langle \mathbb{r}_i, r_j \rangle$ where r is a region in the precise space. Further let us assume that $r_j \in \{r_j^G, r_j^A\}$ where r_j^A means the surrounding area and r_j^G– area occupied by the Ground object. Then we can define (cp. Vasardani, Stirling and Winter, 2017):

$$F\ at\ G \equiv_{def} (in, r_j^A) \in loc(F)$$
$$F\ exactly\_at\ G \equiv_{def} (in, r_j^G) \in loc(F)$$
$$F\ in\ G \equiv_{def} (in, r_j^G) \in loc(F)$$
$$F\ near\ G \equiv_{def} (in, r_j^A - r_j^G) \in loc(F)$$

The regions seem to be mutually nested in the way that $r_j^G \subseteq r_j^A$.

Finally, another approach to vagueness within a region-based formalism is to apply ***tolerance relations*** to RCC relations (Peters and Wasilewski, 2012).

If $x_1, \dots, x_n \in X$ is set of arbitrary spatial entities, *R* set of relations on *X* containing an extended RCC, and if ξ is a set of tolerance relations on *X*, and $t_1, \dots, t_n \in T$ set of temporal intervals, and $c_1, \dots, c_n \in C$ set of concrete locations in physical space then

$$Eig: \langle R, x_1, \dots, x_n, \xi \times t_i \rangle \rightarrow c_j$$

When substituting $x_1, \dots, x_n$ with $o_1, \dots, o_n$ we come to a somewhat similar picture to that of anchoring.

7. Representing motion and change of location

Eigenplace can be also modelled when motion is modelled (cp. Lawvere and Schanuel, 2009, 3f.):

$$f_{motion}: time \rightarrow space$$

or in more detail we can distinguish between

$$f_1: time \rightarrow space$$
$$f_2: space \rightarrow line$$
$$f_3: space \rightarrow plane$$

According to the composition of the functions we can write:

$$f_a: time \rightarrow line$$
$$f_b: time \rightarrow plane$$

where line is, e.g., the level of flight and plane is the place occupied by the shadow of a flying object or position of an object located on earth. If for the sake of simplicity we are assuming that *usual* objects are not flying and this feature is left out of consideration for a while, we can write

$$Eig_{motion}: \langle R_1(o1,..,o_n) \times t_1, \dots, R_n(o1,..,o_n) \times t_n \rangle \rightarrow r_1, \dots, r_n$$

If *o* is an object and *r* is a region (or another object) and $\langle P_1^a, \dots, P_{n:n\geq 1}^a \rangle$ are consecutive segments of a path P^a, then canonically ***entering a region*** can be modeled as a movement at least with *EC*, *PO*, and *TPP* (for details and additional relations cp. Galton, 2000, 282-284):

$$Mov_{Enteringaregion} \langle EC(o,r,P_1^a), PO(o,r,P_2^a), TPP(o,r,P_3^a) \rangle$$

A crucial component of the process of entering a region is the following regularity

$$Enter(o,r,P_1^a) \rightarrow PO(o,r,P_2^a)$$

where *Enter(o,r)* denotes relation of *o* entering *r*. Informally, when an object enters a region, a part of it is inside and a part is outside of that region (i.e., at least in a certain interval of time the relation between *o* and *r* is *PO*).

Canonical set of possibilities before, during entering, and after entering:

$$\langle \Big(\big(DC(o,r) \vee EC(o,r) \vee PO(o,r) \big) P_1^a \Big), EC(o,r,P_2^a), PO(o,r,P_3^a),$$
$$TPP(o,r,P_4^a), \Big(\big(NTPP(o,r) \vee TPP(o,r) \vee PO(o,r) \big) P_5^a \Big) \rangle$$

Another crucial change of location relationship is ***coming into contact***. This can minimally be modelled with *DC* and *EC*:

$$Mov_{coming\ into\ contact} \langle DC(o,r,P_1^a), EC(o,r,P_2^a) \rangle$$

If *o* further enters into *r*, then relation *PO* follows, but this is not necessary the case:

$$\langle DC(o,r,P_1^a), EC(o,r,P_2^a), ((PO(o,r) \vee EC(o,r) \vee DC(o,r))P_3^a) \rangle$$

Another possibility is ***movement of an object from one region to another*** (Galton, 2000, 286). Two possible cases can be distinguished:

a. Movement starts in the first region and ends in the second (e.g., 'John went from his office to canteen');
b. Starting of the movement contains adjacency and ends with adjacency (e.g. 'Mike went from the chair to the window').

Accordingly:

$$Mov_{from\_to\_containing\_region} \langle TPP(o,r_1,P_1^a), TPP(o,r_2,P_2^a) \rangle$$
$$Mov_{from\_to\_adjacent\_region} \langle EC(o,r_1,P_1^a), EC(o,r_2,P_2^a) \rangle$$

Of course, a movement from one region to another can have the starting point as

containing a region and can end with the relation of adjacency (or vice versa).

According to a more recent account (Mador-Haim and Winter, 2015), Eigenplace could be expressed (in a slightly modified way according to the current terminology): a binary relation $\mathbf{far\_from}(\boldsymbol{F}, \boldsymbol{G})$ refers to the following two-place predicate location linking locations (i.e., Eigenplaces) of Figure and Ground – $loc(\boldsymbol{F})$ and $loc(\boldsymbol{G})$. According to Mador-Haim and Winter, Eigenspace (in their terminology) of a Figure is a point (*F*) whereas Eigenspace of a Ground is a region (*G*) (cp. Mador-Haim and Winter, 2015, 442):

$$loc(\boldsymbol{F}) = F$$
$$loc(\boldsymbol{G}) = G$$

This means that the logical form $\mathbf{far\_from}(\boldsymbol{F}, \boldsymbol{G})$ expresses the relation far_from between a point *F* and a region *G*. However, in the current framework this simply means that *F* is a concrete and constrained region whereas *G* is a larger and possibly (although not always) vague or indefinite region (which is the case in far_from).

The core idea by Mador-Haim and Winter (2015) is the Property-Eigenspace Hypothesis (442-443), according to which a relation is between an entity (Figure) and a property (Ground): If $\boldsymbol{F}$ is a Figure and gp a property of the Ground then $\text{far\_from}(loc(\boldsymbol{F}), loc(\text{gp}))$ is far_from relation holding between Eigenplace of Figure and Eigenplace of the properties occupied by Ground. Property-Eigenspace Hypothesis means that every Ground, i.e., property's Eigenspace, "is the union of Eigenspaces for entities in its extension" (Mador-Haim and Winter, 2015, 443). If **gp** is the set of Eigenspaces for properties, then

$$\text{loc(gp)} = \bigcup\{loc(x): x \in \mathbf{gp}\}$$

Therefore,

$$\text{far\_from}(F, \bigcup\{loc(x): x \in \mathbf{gp}\})$$

If *F* is a figure and *G* is a Ground and $\mathcal{G}$ is a set of Grounds, then in the framework by Mador-Haim and Winter (2015, 468) some of the core spatial relations can be modelled

$$\text{far\_from}(F, \bigcup \mathcal{G}) \Leftrightarrow \forall G \in \mathcal{G}.\, \text{far\_from}(F, G)$$
$$\text{close\_to}(F, \bigcup \mathcal{G}) \Leftrightarrow \exists G \in \mathcal{G}.\, \text{close\_to}(F, G)$$
$$\text{outside}(F, \bigcup \mathcal{G}) \Leftrightarrow \forall G \in \mathcal{G}.\, \text{outside}(F, G)$$
$$\text{inside}(F, \bigcup \mathcal{G}) \Leftrightarrow \exists G \in \mathcal{G}.\, \text{inside}(F, G)$$

Consistently with their approach (Mador-Haim and Winter, 2015, 472f.) we can model part-whole relations: if F is a subpart of G, then $loc(F) \subseteq loc(G)$. Therefore, if the regions or elements in the set $\mathcal{F}$ are subparts of *G*, then $\bigcup_{F \in \mathcal{F}} loc(F) \subseteq loc(G)$.
According to our framework

$$\langle far\_from, F, G \rangle \times t_i \rightarrow r_n.$$

8. Conclusion

The Eigenplace relation covers the core of the processes occurring when mapping relational spatial and temporal information to a coordinate space. This mapping is an essential step once cognitive structures (operating in relational spatio-temporal space)

are linked with mathematical coordinate structures operating in numerical terms outside of the human mind.

In our approach, we have defined an ontology that can be used in applications of Eigenplace to resolve spatial vagueness in static and dynamic terms (covering simple and more complex types of movement and motion in space). This corresponds to the idea that the trajectory of an object in space is always linked to a function in time (i.e., there are no spatial movements lacking temporal correlates).

A particularly important direction in our approach is to map vague relational space and accurate space by using the spatial anchoring relation (Galton and Hood, 2005, Hood and Galton, 2006). The anchoring relation is central in spatial communication in general and spatial dialogue systems in particular. Although, cognitively, spatio-temporal existence of objects is always relational and can be cognitively represented in vague or uncertain ways, in virtue of anchoring they can be mapped toprecise coordinate space (i.e., relational objects have exact numerical coordinate correlates), in principle independently of whether we know them or not.

Our developed spatio-temporal ontology operates in an extended RCC formalism (Cohn et al., 1997) and is flexible and open to potentially include other constituents and operators (for functional extensions see also Šķilters et al., 2024). Based on the operator of connectedness (a core operator from which the majority of other operators can be derived) we are able to describe most of the geometrically, topologically, and functionally crucial operators that operate in everyday environments. The most important is the functional operator of locational control, binding the figure and ground object according to the principle that, once the ground is moved in space / time, the figure is moved as well. In these cases, containment is perceived even if it does not apply in the topological sense.

An underlying principle in our approach is the functional prominence of spatio-temporal objects assuming the asymmetry of central (Figure) object and reference (Ground) object that operates in spatial, temporal, and spatio-temporal settings. In the case of temporal situations we are dealing with events as the objects.

Our results can be applied for natural language contexts (especially for modeling the semantics of spatial expressions) but are also usable for non-linguistic spatial information. Although some parts of our approach have been experimentally tested (e.g., Žilinskaitė-Šinkūnienė et al., 2019, Zariņa et al., 2023), there are several spatio-temporal relations (e.g., type of movement and motion, topological features of temporal objects) that can still be both experimentally and computationally tested.

Abbreviations

F – Figure object
G – Ground object
RCC-8 – Region Connection Calculus 8

Acknowledgments

This research was supported by the University of Latvia Foundation and the European Regional Development Fund (ERDF) for postdoc projects (grant agreement no. 1.1.1.2/VIAA/3/19/506). A part of this work (Jurģis Šķilters) was supported by the Fulbright Scholar Program (2013/2014).

References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 832-843.

Asbury, A., Gehrke, B., Van Riemsdijk, H., Zwarts, J. (2008). Introduction: Syntax and semantics of spatial P. In Asbury, A., Dotlačil, J., Gehrke, B., Nouwen, R. (Eds.). *Syntax and semantics of spatial P* (pp. 1- 32). Amsterdam: John Benjamins.

Bennett, B., Düntsch, I. (2007). Axioms, algebras and topology. In *Handbook of spatial logics* (pp. 99-159). Dordrecht: Springer Netherlands.

Bennett, B., Galton, A. P. (2004). A unifying semantics for time and events. *Artificial Intelligence*, 153(1), 13-48.

Bittner, T., Stell, J. G. (2000). Approximate qualitative spatial reasoning. *Spatial Cognition and Computation*, 2(4), 435-466.

Chen, H., Vasardani, M., Winter, S. (2017). Geo-referencing Place from Everyday Natural Language Descriptions. *arXiv preprint* arXiv:1710.03346.

Carlson, L., Covell, E. (2005). Defining functional features for spatial language. In Carlson, L., Van der Zee, E. (Eds.). *Functional features in language and space: insights from perception, categorization, and development* (pp.175-190). Oxford: Oxford University Press.

Clarke, B. L. (1981). A calculus of individuals based on "connection". *Notre Dame Journal of formal logic*, 22(3), 204-218.

Clarke, B. L. (1985). Individuals and points. *Notre Dame Journal of Formal Logic*, 26(1), 61-75.

Cohn, A. G. (1995). Qualitative shape representation using connection and convex hulls. *Proceedings of Time, Space and Movement: Meaning and Knowledge in the Sensible World*, 3-16.

Cohn, A. G., Bennett, B., Gooday, J., Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1, 275-316.

Cohn, A. G., N. M. Gotts, Z. Cui, D. A. Randell, B. Bennett, M. Gooday. (1998). Exploiting temporal continuity in qualitative spatial calculi. In Egenhofer, M.J., Colledge, R. G. (Eds.), *Spatial and temporal reasoning in GIS* (pp. 5-24). New York, New York: Oxford University Press.

Cohn, A. G., Li, S., Liu, W., Renz, J. (2014). Reasoning about topological and cardinal direction relations between 2-dimensional spatial objects. *Journal of Artificial Intelligence Research*, 51, 493-532.

Cohn, A. G., Randell, D. A., Cui, Z. (1995). Taxonomies of logically defined qualitative spatial relations. *International journal of human-computer studies*, 43(5-6), 831-846.

Cohn, A. G., Renz, J. (2008). Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3, 551-596.

Cohn, A. G., Varzi, A. C. (2003). Mereotopological connection. *Journal of Philosophical Logic*, 32, 357-390.

Davidson, D. (1969). *The individuation of events* (pp. 216-234). Springer Netherlands.

Davis, E., Marcus, G., Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial intelligence*, 248, 46-84.

De Laguna, T. (1922). Point, line, and surface, as sets of solids. *The Journal of Philosophy*, 449-461.

Dong, T. (2008). A comment on rcc: From rcc to rcc++. *Journal of Philosophical Logic*, 37(4), 319-352.

Dube, M. P. (2017). Topological augmentation: A step forward for qualitative partition reasoning. *Journal of Spatial Information Scienc*e, 14, 1-29.

Düntsch, I., Wang, H., McCloskey, S. (2001). A relation–algebraic approach to the region connection calculus. *Theoretical Computer Science*, 255(1-2), 63-83.

Gardenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT press.

Galton, A. (2000). *Qualitative spatial change*. New York: Oxford University Press.

Galton, A. (2004). Fields and objects in space, time, and space-time. *Spatial cognition and computation*, 4(1), 39-68.

Galton, A. (2009). Spatial and temporal knowledge representation. *Earth Science Informatics*, 2, 169-187.

Galton, A. (2014). Discrete mereotopology. In *Mereology and the Sciences: Parts and Wholes in the Contemporary Scientific Context* (pp. 293-321). Cham: Springer International Publishing.

Galton, A., Hood, J. (2005). Anchoring: a new approach to handling indeterminate location in GIS. In A.G. Cohn, D.G. Mark (Eds.). *International Conference on Spatial Information Theory; Lecture Notes in Computer Science*, Volume 3693 (pp. 1-13). Berlin: Springer.

Gerla, G. (1995). Pointless geometries. In *Handbook of incidence geometry* (pp. 1015-1031). North-Holland.

Gerevini, A., Renz, J. (2002). Combining topological and size information for spatial reasoning. *Artificial Intelligence*, 137(1-2), 1-42.

Golledge, R. G. (1992). Place recognition and wayfinding: Making sense of space. *Geoforum*, 23(2), 199-214.

Hood, J. (2007) Taking Location Seriously: Is Location a Function or a Relation? In Probst, F., Kessler, C. (Eds.). GI-Days 2007 – *Young Researchers Forum*. IfGIprints 30 (pp. 211-215). ISBN: 978-3-936616-48-4.

Hood, J., Galton, A. (2006). Implementing anchoring. In Geographic Information Science: *4th International Conference, GIScience 2006, Münster, Germany, September 20-23, 2006. Proceedings 4* (pp. 168-185). Springer Berlin Heidelberg.

Jiang, J., Worboys, M. (2009). Event-based topology for dynamic planar areal objects. *International Journal of Geographical Information Science*, 23(1), 33-60.

Kim, J. (1973). Causation, nomic subsumption, and the concept of event. *The Journal of Philosophy*, 70(8), 217-236.

Kontchakov, R., Pratt-Hartmann, I., Zakharyaschev, M. (2010). *Interpreting Topological Logics over Euclidean Spaces*. In Lin, F., Sattler, U., Truszczynski, M. (Eds.), *Proc. of the 12th International Conference on Principles of Knowledge Representation and Reasoning (KR 2010)* (pp. 534-544). AAAI Press

Kracht, M. (2002). On the semantics of locatives. Linguistics and *Philosophy* 25, 157– 232.

Kracht, M. (2008). The fine structure of spatial expression. In Asbury, A., Dotlačil, J., Gehrke, B., Nouwen, R. (Eds.). *Syntax and semantics of spatial P* (pp. 35-62). Amsterdam: John Benjamins.

Landau, B. (1996). Multiple geometric representations of objects in languages and language learners. In Bloom, P., Peterson, M.A. (Eds.), *Language and Space. Language, Speech, and Communication* (pp. 317–63). Cambridge, MA: MIT Press

Lawvere, F. W., Schanuel, S. H. (2009). *Conceptual mathematics: a first introduction to categories*. 2nd Edition. Cambridge: Cambridge University Press.

Lewin, K. (1936). *Principles of topological relations*. New York and London: McGraw-Hill. DOI, 10, 10019-000.

Li, S., Cohn, A. G. (2012). Reasoning with topological and directional spatial information. *Computational Intelligence*, 28(4), 579-616.

Lyons, J. (1977). *Semantics: Volume 2* (Vol. 2). Cambridge University Press.

Mani, I., Pustejovsky, J. (2012). *Interpreting motion: Grounded representations for spatial language* (No. 5). Oxford University Press.

Mador-Haim, S., Winter, Y. (2015). Far from obvious: the semantics of locative indefinites. *Linguistics and Philosophy*, 38(5), 437-476.

Miller, G. A., Johnson-Laird, P. N. (2013). *Language and perception. In Language and Perception*. Harvard University Press.

Peters, J. F., Wasilewski, P. (2012). Tolerance spaces: Origins, theoretical aspects and applications. *Information Sciences*, 195, 211-225.

Piñón, C. J. (1993). Paths and their names. In *Chicago Linguistics Society* (Vol. 29, No. 2, pp. 287-303).

Pustejovsky, J. (2013). Where Things Happen: On the Semantics of Event Localization. In *Proceedings of ISA-9: International Workshop on Semantic Annotation*.

Randell, D. A., Cui, Z., Cohn, A. G. (1992). A spatial logic based on regions and connection. *Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo* (pp. 165–176).

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, 80(1-2), 1-46.

Svenonius, P. (2010). Spatial p in English. Mapping spatial PPs: *The cartography of syntactic structures*, 6, 127-160.

Stell, J. G. (2000). Boolean connection algebras: A new approach to the Region-Connection Calculus. *Artificial Intelligence*, 122(1-2), 111-136.

Šķilters, J., Zariņa, L., Glanzberg, M. (2024). Towards a Framework for Functional Representation of Spatial Relations. Baltic Journal of Modern Computing, *12*, (1), 15-29.

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 2). Cambridge, MA: MIT press.

Tomko, M., Winter, S. (2013). Describing the functional spatial structure of urban environments. *Computers, Environment and Urban Systems*, 41, 177-187.

Van Lambalgen, M., Hamm, F. (2005). *The proper treatment of events*. Malden, MA: Blackwell.

Vasardani, M., Stirling, L. F., Winter, S. (2017). The preposition at from a spatial language, cognition, and information systems perspective. *Semantics and Pragmatics*, 10.

Wolter, F., Zakharyaschev, M. (2000). Spatial representation and reasoning in RCC-8 with Boolean region terms. In *Proceedings of the 14th European Conference on Artificial Intelligence* (pp. 244-248).

Wunderlich, D. (1991). How do prepositional phrases fit into compositional syntax and semantics? *Linguistics*, 29, 591-621.

Wunderlich, D. (1993). On German um: semantic and conceptual aspects. *Linguistics*, 31, 111-133.

Wunderlich, D., Herweg, M. (1991). Lokale und Direktionale. In von Stechow, A., Wunderlich, D. (Eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung* (pp. 758-785). Berlin: De Gruyter.

Zariņa, L., Šķilters, J., Draudiņš, M., Žilinskaitė-Šinkūnienė, E. (2023). Impact of Scale on the Perception of Proximity as Represented in Latvian. *Baltic Journal of Modern Computing*, *11*(4), 523-541.

Zwarts, J. (1997). Vectors as relative positions: a compositional semantics of modified PPs. *Journal of Semantics*, 14(1), 57-86.

Zwarts, J. (2017). Spatial semantics: Modeling the meaning of prepositions. *Language and Linguistics Compass*, 11(5), e12241.

Zwarts, J., Winter, Y. (2000). Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information*, 9, 169-211.

Žilinskaitė-Šinkūnienė, E., Šķilters, J., Zariņa, L. (2019). Containment and support in Baltic languages: Overview, experimental evidence, and an extended RCC as applied to Latvian and Lithuanian. *Baltic Journal of Modern Computing*, *7*(2), 224-254.

Received June 6, 2023, revised July 15, 2024, accepted July 22, 2024

Baltic J. Modern Computing, Vol. 12 (2024), No.3, 248-258
https://doi.org/10.22364/bjmc.2024.12.3.02

Digital Learning Model

Sarma Cakula

Vidzeme University of Applied Sciences
Tērbatas iela 10, Valmiera, LV - 4202, Latvija

sarma.cakula@va.lv

ORCID 0000-0003-4831-2057

Abstract. Distance learning is becoming increasingly important today, it is essential to identify and explore digital learning opportunities, developing technological support and digital learning methods accordingly. One of the most important aspects of e-learning is the personal motivation of the student, so the learning process must involve the student in an active way. Additionally, technologies should be such as to support the increase of this motivation. There is a growing shift to using active learning methods in full-time study. Various e-learning platforms have been developed as more and more researchers are exploring the development of digital teaching and learning methodologies. However, there is currently no established technological framework to support the various active digital learning methods in a remote study environment. The aim of the paper is to develop and evaluate a conceptual technological model of active digital teaching and learning. Theoretical and statistical methods are used to reach this aim. The result of the paper is a technological model of e-teaching and e-learning comprising several interconnected parts of a system that promotes the active involvement of both the student and the teacher in the learning process, ensuring a higher quality knowledge sharing between both sides.

Keywords: active learning methods, e-learning, embedded systems, digital technological model.

1. Introduction

Changes in the global economy and politics are accelerating, with education, research and culture all being affected. Work equipment is becoming more sophisticated and complex, requiring more and more knowledge and skills from the workers. This leads to recognizing that knowledge is becoming increasingly valuable. The digital transformation of the global economy and society, and the speed of change is increasing the complexity of today's world as it becomes more connected and better educated. This complexity and speed of change mean that connecting education to the trends shaping the world we live in is now increasingly urgent. The information technology sector is developing rapidly in all areas of the economy but is lagging in the growing need to use it in a specific direction. The acquisition of new knowledge and skills has been seen as an important aspect of university education, signifying an increase in the importance of studies, both in full-time programmes and in individual courses. Today's education system is increasingly moving towards active learning, which includes active teaching methods, but digital tools and models that fully support active learning are still lacking. There are difficulties in

replicating traditional teaching methods resulting in unpredictable learning outcomes. Modern information systems and technologies support continuous development of digital tools, opening opportunities for strengthening the education system.

At a time when the need for distance learning is becoming increasingly important, it is essential to identify and explore digital learning opportunities, developing technological support and digital learning methods accordingly. One of the most important aspects of e-learning is the personal motivation of the student, so the learning process must involve the student in an active way. There is an increasing tendency to use active learning methods in full-time study. Various e-learning platforms have been developed as more and more researchers are exploring the development of digital teaching and learning methodologies. However, there is currently no established digital learning technological framework to support the various active learning methods.

The problem is the lack of active learning methods and knowledge management technologies in the study process. **The aim of the paper** is to develop a conceptual technological model of active digital teaching and learning.

Theoretical-technological model produces the event structure that is possible from the point of view of pragmatic competence and technology implements the event structure in accordance with pragmatic performance in order to realize surface uniqueness. The theoretical and the technological components of the model share at least three essential features: First of all, they follow the principle of modularity along the same kinds of modules, secondly, they adhere to the duality of competence and performance, thirdly, they handle the multimodal nature of real human-to-human (and, ultimately, human-to-machine) communication (Hunyadi, 2011).

Research question-what technological model improves student's knowledge in learning process. Research tasks includes developing of theoretical digital learning model, experiment of using digital active learning methods and evaluating of main results.

The paper examines the feasibility of applying digital active learning methods in a digital environment, assesses the need to personalize the study process and establishes a common technological framework for the use of active learning methods. The paper develops a conceptual model for an e-learning technological platform to enable the use of active learning and teaching methods in studies at higher education institutions.

2. Main focus of today's digital learning

2.1. Research background

Lifelong learning is one of the most important parts of today's education system, requiring the development of prior learning and professional skills in line with the requirements of the profession. Learning at a distance in a digital environment is becoming particularly important. Strategically, the main goal of sustainable national development and human well-being is to focus on education for personal development.

Historical circumstances have changed, the content of education has changed, new educational paradigms have emerged, and thus both the professional competence of the teacher and the use of technology have acquired new significance. The supply of information is increasing, and it is becoming more and more important to be able to navigate quickly and efficiently through large amounts of accessible data, to acquire new skills and competences.

Digital learning technologies support the digitization of the learning experience and facilitate online mobility and include any communication, information and technological tools that contribute to improved teaching, development, and assessment.

Digital learning technologies include:

- Mobile learning and apps
- Gamification
- Virtual classrooms
- Artificial intelligence (AI**)**
- Lifelong learning technologies
- Immersive learning technologies
- Nano-learning technologies (Cumraeg, 2022).

These technologies are developed by various EdTech companies, including the World Bank Group, based on scientific research in this field (The World Bank, 2022). EdTech companies have also been either a help or a hindrance. Technology has been positioned as the solution to pedagogic innovation, when the reality is that learners need to take responsibility for their learning in order for sustained progress to be made (Learnlife, 2022). There is some research into how digital learning has been affected by Covid pandemic. There is a growing need to advance digital education ecosystems and technologies (Mihovska et.al., 2021). All over the world, both universities and moodle create and organize a wide variety of courses in many languages, but they need to be adapted to each group of students.

Alamri et. al. provide an overview of personalized learning theory and learning technology that supports the personalization of higher education. They have analized three technological models that support personalized learning within various learning environments in higher education. Personalized learning is "an educational approach that tailors learning around each individual student's needs, interests, and abilities. Each student is given differentiated instruction based on their personal learning characteristics" (Raudys, 2021). However, they emphasize the lack of data-driven and independent research studies that could enable increased effectiveness and impact of the personalized learning and technology models on student learning (Alamri et.al., 2021). The depth of teaching properties of digital resources in guidance is discussed through the possibility of identifying orientation models in their theoretical structure (Payo et.al., 2013). It is possible to use the help of artificial intelligence to choose your own learning path (Cognitive Class.ai, 2024). Some authors describe the possibility to integrate a knowledge assessment system based on concept maps with a personalized study planning prototype and examine its use in personal study planning (Rollande et.al., 2017).

There is still a lack of extensive research into personalized digital learning technological models that focus on increasing a student's motivation.

2.2. Digital active learning methods

The paper focuses on technological solutions that provide educational methods which enhance everyone's ability to acquire knowledge, values and skills needed to participate in decision-making for individual or collective action at local and global levels to improve the quality of life without compromising the needs of future generations. Learning materials and methods in a digital environment enable the rapid and secure introduction of new knowledge and the mutually beneficial exchange of data and knowledge, which are key to sustainable education. Active learning methods develop learner's ability to react

flexibly in a competitive environment but have so far been used mainly in face-to-face studies.

Mercat Christian presents Active Learning Methodology, surveying its history, main existing tools and supporting evidence, with an emphasis on mathematics and higher education, in particular engineering (Mercat, 2021).

Student motivation, engagement and interest in their own learning are imperative for a successful and student-centred education. The global education trend has shifted to a clearer focus on '21st century skills' or transversal competences. Humanisation, accessibility, openness, and diversity of the educational environment are the guarantors of sustainable development of education. Several scholars have provided theoretical justification for distance learning ideas related to the expansion of the opportunities offered in the context of home-based education and international or cross-border education (Katane et.al., 2012). We need to develop a different orientation when thinking about new technologies in education - not just as tools or delivery systems, but as a set of resources and capabilities that enable us to rethink our educational goals, methods, and institutions (Burbules et.al., 2020). Active learning methods provide new content created by the teacher or student (Kim et.al., 2012). Technology-enhanced learning environments create flexibility and sustainability in education (Cakula, 2018).

The author carried out an experiment at Vidzeme University of Applied Sciences (VIA) that involved experiential learning and active participation based on collective coding exercises (VIA student codes), quizzes, projects, and other approaches. Experiential learning in algorithms and statistics courses took the form of practical exercises in the development of collective solutions. Work on algorithms and statistics exercises was carried out in small groups, with regular feedback data collected in a number of ways to serve as input to the knowledge discovery process to support active learning later on (Cakula, 2021). The methods used were various active learning methods such as Dotmocracy, Fishbowl, Survey, Index Card Pass, Flipped Classroom, Complete Turn Taking, Respond, React, Reply, Round Table, Think-Pair-Share, Post It Parade, including also solving different exercises on an algorithm theory (Hattie et.al.,2007).

2.3. Personalized learning process

Instruction tailored to the unique pace of different students is known as personalized learning. Personalized learning is a method of teaching in which the content, technology and pace of learning are based on the abilities and interests of each learner.

There are five steps to personalizing learning:

- set clear and specific goals,
- make goals challenging and realistic,
- make goals dynamic and review them regularly,
- let learners know their progress,
- involve supporters (parents, friends, etc.) (Rogers, 1997).

In this case, the academic goals remain the same for the group of students, but individual students can progress through the curriculum at different speeds based on their specific learning needs. This is particularly true in e-learning, where each student chooses to study at his or her own time, place, and pace. Personalised learning includes adaptive learning, individualized learning, differentiated learning and competency-based learning (Briggs et.al, 2009). Adaptive learning is when technology is used to assign human or digital resources to learners based on their unique needs. Personalized learning states that

the pace of learning is adapted to the needs of individual students. Differentiated learning approach states that learning is adapted to the needs of individual students. Competency-based learning provides learners with the opportunity to progress through a learning pathway based on their ability to demonstrate competence, including the application and creation of knowledge, as well as skills and dispositions. Academic objectives, curriculum, and content, as well as method and pace may vary in a personalised learning environment. Unlike individualized learning, personalized learning involves students in the design of learning activities and is based more on the student's personal interests and motivation to acquire knowledge and skills.

Individual perception can be classified as a form of nomothetic psychology and is developed by Socionics' theory based on Jungian four personality types. Jung mostly focussed on personality types as individuals. Meanwhile, Socionics states that there are 16 types of personalities and respectively 16 types of possible perception of information. All of people have their strengths and weaknesses and Socionics has defined what the strengths and weaknesses of each sociotype are in perception of information (Desmarais, 2006; Grant, 2014; Sampson et.al., 2002). To ensure the highest quality e-learning, the individual characteristics of learners must be considered. It should be noted that this is one of the biggest advantages of e-learning, because unlike a standard learning environment where students listen to a lecture together and do the same homework and tests, e-learning can provide tailored information based on the student's most pronounced perceptual channel. In the literature, these are also referred to as modalities, Fleming's VARK model or simply as perceptions. Four perceptual channels are distinguished:

- auditory
- visual
- reading, writing
- kinaesthetic (Othman et.al., 2010).

Building on all these aspects, smart learning environments provide students with adaptive and personalized learning and assessment, including multimodal/multisensory interaction technologies and advanced interfaces. An industry-driven approach in collaboration with academics will lead to market-oriented education. New learning individually oriented methodology should be developed based on individual human perception – how individuals select and process information - Neil Fleming's pedagogic theory, educational psychology, and artificial intelligence for formal and informal education, including workplace learning. Increasingly, the course leader collaborates with students in their studies, which promotes faster and more effective knowledge sharing and the creation of new knowledge.

3. Digital learning framework

3.1. Participants

Research base is 348 students from Information technology and Business Management bachelor programs in Vidzeme University of Applied Sciences, who participated in the experiment learning course "Statistics" between the years of 2014 and 2023. 89% of them chose to use the Learning Support System.

3.2. Material

The basis was the Moodle platform, which integrated various modules. Information Technology students took 6 ECTS Statistics course, while Business Management students took 3 ECTS Statistics course. Boths groups got 2nd year bachelor program course according to business management and information technology accredited study programs in Latvia.

At the beginning of the experiment, the basic content and materials were prepared by the lecturer, but the students had the opportunity to prepare the course content of each meeting in various forms under the guidance of the teacher - both text documents, using images, colors, video, audio, multimedia, etc. according to their sociotype. These materials were placed in the moodle system for use in the following years of studies. A database of various materials was formed, where it was expanded in each subsequent year. Students determined their sociotype using Jung's short test consisting of 60 questions (Similarminds, 2024).

3.3. Design

The digital learning framework is built on advances in neuroscience, pedagogical and learning theories, educational psychology as well as artificial intelligence including modal/multisensory technologies. It provides gaining access from both sides - teacher and student (Fig. 1). Framework includes 4 main parts: course content, active learning environment, learning support system and feedback to both - student and teacher. Course's content foundation is developed by the teacher and may be supplemented by student-generated content under the guidance of the teacher. It is very important that students take part in developing content – this is one of the most powerful methods of enabling students to look at things from a teacher's perspective. It develops a deeper understanding of necessary knowledge and skills. Active learning environment includes three main parts – possibility to build a course map choosing between different active learning methods and using tools embedded in the learning environment for using active methods in both group and individual work (Sampson et.al., 2022).

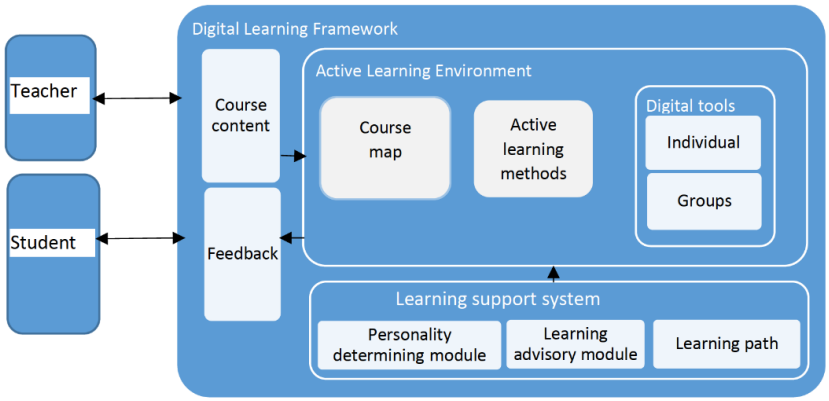


Fig. 1. Digital Learning Conceptual Model

Learning support system includes personality determining module (Cakula, 2018), learning advisory module (Cakula et.al, 2019) and learning path (Nabizadeh et.al, 2020) that can be developed for every individual student. Course content in the beginning is created by the teacher and later developed by students using course map and active learning methods. In the leaning process personality determining module, learning advisory module, and learning path could be accessed depending on student's choice.

Personality determining module will profile users, assessing their personality based on Socionics theory, Neil Fleming's pedagogic theory, educational psychology, and artificial intelligence to provide adaptive learning content for improving performance in learning. Learning advisory module system will manage the learning path for each learner through the course content, using different course units, access from different devices and evaluation feedback. As a result, it will be able to advise the learners to follow a different learning design (relevant to their personality type) or access different learning resources (relevant to their information processing preferences). The tutor / administrator will also be informed for the learners' progress during each module / course.

Learning path can be created based on an algorithm developed by scientists in Chine where they have designed a multidimensional knowledge graph framework that separately stores learning objects organized in several classes and proposes six main semantic relationships between learning objects in the knowledge graph. Learning path recommendation model is designed for satisfying different learning needs based on the multidimensional knowledge graph framework, which can generate and recommend customized learning paths according to the e-learner's target learning object (Daqian, 2020).

Digital Learning Environment, for example Moodle, is the main system where access from teacher and student will start. The course is divided in several learning units. Course content is organized using questions in the beginning of every learning unit. This will offer next steps for each student based on their choice in using Learning Support System. Both – Learning Support System and Active Learning Environment - are embedded systems included in the main Digital Learning Environment. Digital Tools are accessed by links from every learning module. In the future there should be further research into embedding this in the main Digital Learning Environment Digital Tools system.

3.4. Procedure (Research Process).

At the beginning of the course, each student was offered the opportunity to study in a traditional way or use the opportunity to determine his sociotype and use the approach of the system mentioned in the article. 85% of the students were full-time students, and 15% were e-study students. Those students who chose to use the study materials offered to the sociotype (87% on average) could, respectively, voluntarily use the learning advisory module and learning path module integrated in moodle. All students took part in active Learning environment managed by teacher.

Evaluation for each student took place in accordance with the accredited and approved course description by the teacher using tests and activity in the course. All students took all tests included in the course.

Questionnaires and interviews were performed at the end of the course to obtain students' views on learning process.

4. Main research results

There is a correlation between personality type and the learning program – Information Technology students were mostly found to be Introvert and Judging (characteristics defined by Socionics) but Business Management students covered all personality types in every study year.

Depending on the personality type students took part in the developing learning materials for every course content module. There were different evaluation methods in every course content unit and the final grade constantly improved, starting with an average grade of 6,34 in 2014 and reaching 8,14 in 2023 (Fig. 2).

Kalmogorov-Smirnov nonparametric tests were used for testing normal distribution (Fig. 3).

H0 - final grade for IT students in Statistics course follows normal distribution is accepted on probability level 95%.

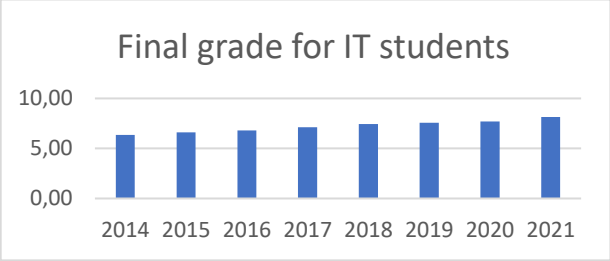


Fig. 2. Final grade for Information technology students in Statistics course

		Final grade of IT students
N		**214**
Normal Parameters	*Mean*	7,22
	Std. Deviation	1,52
Most Extreme Differences	*Absolute*	,05
	Positive	,03
	Negative	-,05
Kolmogorov-Smirnov Z		,73
Asymp. Sig. (2-tailed)		,658

Fig. 3. Kalmogorov-Smirnov nonparametric test for normal distribution

From 2018 in addition active digital learning methods and course map was included. Until 2018, the average results for IT students improved by 6.9% per year, but in 2018 and

after the results improved in average by 7.5%. Different situation is for business students - there were no regular progress from year to year (Fig. 4).

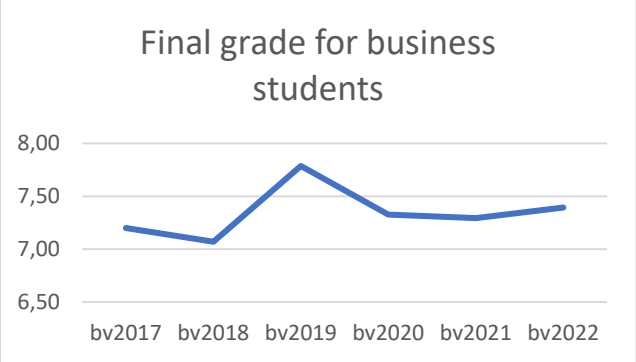


Fig. 4. Final grade for business students in Statistics course

Active digital learning methods positively influence results but there is no clear trend for growing results. In discussion students accepted that this solution is much better for them than tradition learning, and they were very interested to understand their sociotype.

5. Conclusion

Today, the pedagogical paradigm is shifting from teaching to learning and, by extension, focuses on what the student requires from the teacher. Students and their activities have a great influence on teaching methods, content and technological tools used. A technologically supported information system is necessary to ensure the delivery of relevant content and the promotion of a coherent study system. Furthermore, it can motivate students to delve deeper in the courses offered and to acquire the competences needed in the labour market. In a quickly evolving information society it is increasingly important to deliver the right information to the right learner, quickly.

The working environment is changing with the rapid development of technology and knowledge, so it is important to keep abreast of changing market conditions and not only apply available technological solutions but also develop new applications of technology in areas of societal need to contribute to the overall growth of the economy. The development of a sustainable society is influenced by the variety of methods and technologies available.

The level of students' motivation to learn is becoming more and more important. In the digital environment, there is much less support from other students and from the lecturer, so the digital environment should be one that includes personal support for each individual student and offers learning according to the way each one perceives information.

The technological digital learning model created draws on advances in learning theory, neuroscience, artificial intelligence, educational psychology, and modal/multisensory technology. It allows for teachers and students to cooperate in a common system using different embedded tools supporting active learning methods.

References

Alamri, H. A., Watson, S., Watson, W. (2021). Learning Technology Models that Support Personalization within Blended Learning Environments, In: *Higher Education*, TechTrends, **65**(1), 62-78.

Briggs Myers, I., McCaulley, H. M., Quenk, L. N., Hammer, L. A., Wayne, D. M. (2009). MBTI Step III Manual: Exploring Personality Development Using the Myers-Briggs Type Indicator Instrument, *Consulting Psychologists Press*.

Burbules N.C., Fan G., Repp P**.** (2020). Five trends of education and technology in a sustainable future, *Geography and Sustainability*, **1**(2), 93-97.

Cakula, S. (2018). Smart Technological Learning Conceptual model, *International journal of Engineering and Technology*, **7**(2), 152-156.

Cakula, S., Majore, G. (2019). Future Generation Education Technological Model, *Proceedings - 2019 IEEE 9th International Conference on Intelligent Computing and Information Systems, ICICIS 2019*, 9014852, 371-376.

Cakula, S. (2021). Active Learning Methods for Sustainable Education Development, *Rular Environment. Education. Personality*, **14**, 59-65.

Cognitive Class.ai (2024). learning Path. https://cognitiveclass.ai/learn

Cumraeg - North Weles Management Scool (2022). What are digital learning technologies. https://online.glyndwr.ac.uk/what-are-digital-learning-technologies/

Daqian Shi, Ting Wang, Hao Xing and Hao Xu. (2020 May). A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning, *Knowledge-Based Systems*, **195**, 11, 105618.

Desmarais, M.l C., Gagnon M. (2006). Bayesian Student Models Based on Item to Item Knowledge Structures , *EC-TEL 2006: Innovative Approaches for Learning and Knowledge Sharing*, 111-124.

Grant, P., Basye, D. A. (2014) Personalized Learning: Guide for Engaging Students with Technology, *International Society for Technology in Education*. Oregon, Washington.

Hattie, J., Timperley, H. (2007). The Power of Feedback, *Review of Educational Research*, **77**(1), 81-112.

Katane, I., Katans, E., Vavere, G. (2012). Environment of distance learning for humanization and democratization of education: the historical aspect", In V. Dislere (Ed.), *The Proceedings of the International Scientific Conference Rural Environment. Education. Personality (REEP)*, **5**, 35-42.

Kim J., Hwang J., Chi S., Seo J. (2020). Towards database-free vision-based monitoring on construction sites: A deep active learning approach. *Automation in Construction*, **120**, 103376.

Learnlife. (2022). Technology & Digital Learning Platforms. https://www.learnlife.com/learning-paradigm/technology-digital-platforms

Mercat Christian . (2022). Introduction to Active Learning Techniques**.** Open Education Studies **4**(1):161-172.
file:///C:/Users/Admin/Downloads/Introduction_to_Active_Learning_Techniques.pdf

Mihovska, A., Prevedourou, D., Tsankova, J., Manolova, A., Poulkov, V. (2021). Building Adaptive and Inclusive Education Readiness through Digital Technologies, *Joint 6th International Conference on Digital Arts, Media and Technology with 4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, ECTI DAMT and NCON 2021*, **9425728**, 384-388.

Moodle (2024) Making quality online education accessible for all. https://moodle.com/services/learning-design/

Nabizadeh, A. H., Leal, J. P., Hamed N. Rafsanjani, H. N., RatnShah, R. (2020 November). Learning path personalization and recommendation methods: A survey of the state-of-the-art, *Expert Systems with Applications*, **159**, 30, 113596.

Othman, N., Amiruddin, M. H. (2010). Different Perspectives of Learning Styles from VARK Model, *Procedia - Social and Behavioral Sciences*, **7**, 652-660.

Payo, A. R., Martín, S. N., Sánchez, E. M. T. (2013). The technological model in the school guidance into digital educational resources, *ACM International Conference Proceeding Series*, 585-589.

Raudys, J. (2021). Personalized learning strategies to implement in class and examples. https://www.prodigygame.com/main-en/blog/personalized-learning/

Rogers, J. (1997). Sixteen personality types at work in organizations, London, Management Futures.

Rollande, R., Grundspenkis, J. (2017). Personalized Planning of Study Course Structure Using Concept Maps and Their Analysis, *Procedia Computer Science*, **104**, 152-159.

Sampson, D., Karagiannidis, C., Kinshuk (2002). Personalised Learning: Educational, Technological and Standardisation Perspective, *Interactive Educational Multimedia*, *Special Issue on Adaptive Educational Multimedia*.

Similarminds (2024). Free Jung Personality Test (similar to MBTI / Myers Briggs) https://similarminds.com/jung.html

The University of Texas at Austin (TEXAS). (2020). Flipped Classroom. https://facultyinnovate.utexas.edu/instructional-strategies/flipped-classroom

The World Bank. (2022). Digital Technologies in Education. https://www.worldbank.org/en/topic/edutech

Received February 13, 2024, revised June 28, 2024, accepted August 8, 2024

Baltic J. Modern Computing, Vol. 12 (2024), No. 3, pp. 259–269
https://doi.org/10.22364/bjmc.2024.12.3.03

Quantum Algorithm for the Domatic Number Problem

Andris AMBAINIS, Iļja REPKO

Centre for Quantum Computer Science, Faculty of Sciences and Technology, University of Latvia
Raiņa bulvāris 19, Riga, LV–1586, Latvia

andris.ambainis@lu.lv, ilja.repko@gmail.com

ORCID 0000-0002-8716-001X, ORCID 0009-0003-3154-1803

Abstract. In this paper we design a quantum algorithm for the NP-complete problem of finding the domatic number. The DOMATIC NUMBER problem asks to determine the largest integer k, such that a given undirected graph with n vertices can be partitioned into k pairwise disjoint dominating sets. This problem finds significant applications, such as in wireless sensor networks, where the selection of multiple dominating sets balances energy consumption and extends network lifetime. Each node communicates exclusively with designated nodes, and dominant sets ensure network resilience by enabling seamless replacement in case of node cluster failure. The importance of this problem lies in its implications for network optimization, highlighting the advantages of quantum computing in addressing complex combinatorial challenges. We present a quantum algorithm that solves this problem in time $\mathcal{O}(2.4143^n)$, which we further improve to $\mathcal{O}(2.3845^n)$.

Keywords: domatic number, quantum algorithm, dominating set

1 Introduction

In this paper we have discovered two quantum algorithms to find the domatic number $d(G)$ for the graph G. The problem was discovered in (Chang, 1994).

These algorithms solve a critical problem in graph theory that affects diverse fields, including wireless sensor networks (Jiguo, Qingbo, Dongxiao, Congcong and Guanghui, 2014). Efficiently selecting and managing multiple dominating sets in these networks is crucial for saving energy and extending network lifespan. Quantum computing offers a promising approach to enhance network optimization and analyze data structures in new ways. This shows how quantum algorithms can tackle complex problems that traditional computers struggle with.

Previous results. Classically the problem can be solved in time $3^n \cdot n^{\mathcal{O}(1)}$ (Fomin, Kratsch, 2010) using an algorithm similar to Lawler's dynamic programming algorithm (Lawler, 1976). In the first half of 2005, Riege T. and Rothe J. "broke through" the 3^n barrier, solving the DOMATIC NUMBER problem for 3 dominating sets in time $\tilde{\mathcal{O}}(2.9416^n)$ (Riege, Rothe 2005). Later, authors combined the inclusion-exclusion approach with the Fomin algorithm (Fomin, Grandoni, Pyatkin, Stepanov, 2005) to solve the problem in time $\tilde{\mathcal{O}}(2.695^n)$ for 3 dominating sets (Riege, Rothe, Spakowski, Yamamoto, 2007).

Nevertheless the best known classical algorithm was presented by Johan M.M. van Rooij which solved the problem in time $\mathcal{O}(2.7139^n)$ (van Rooij, 2010).

Until recently, no quantum algorithm was known for this problem. Then, two algorithms were developed independently: the algorithm in this paper (which was developed and presented as a bachelor's thesis in June 2023 in University of Latvia (Repko, 2023)) and an algorithm (Gaspers and Li, 2023) which appeared in November 2023. The quantum algorithm of Gaspers and Li is faster, with the complexity of $\mathcal{O}((2-\epsilon)^n)$ but our algorithm is simpler.

Main contributions:

- Firstly, we designed an algorithm that solves the DOMATIC NUMBER problem in time $\mathcal{O}(2.4143^n)$. This technique combines Lawler's algorithm for finding the chromatic number (Lawler, 1976) with Grover's search (Durr and Høyer, 1996) and the naive computation of minimal dominating sets using dynamic programming in time $2^n \cdot n^{\mathcal{O}(1)}$.
- Secondly, we developed an algorithm that utilizes the precomputation of improved minimal dominating sets (Fomin, Grandoni, Pyatkin, and Stepanov, 2008), solving the DOMATIC NUMBER problem in time $\mathcal{O}(2.3845^n)$.
- Thirdly, we provided a data structure that precomputes all minimal dominating sets in time $\mathcal{O}^*(2^n)$ and allows each set to be obtained in time $\mathcal{O}(n^c)$.

The article consists of four sections. The first section is the introduction, where the main problem and contributions are presented. The second section covers the preliminaries, theorems, and techniques important for implementing the algorithm on a quantum computer. The third section is divided into two parts: the first part describes the simple algorithm that solves the problem in time $\mathcal{O}(2.4143^n)$, and the second part presents an improved version of it that solved DOMATIC NUMBER in time $\mathcal{O}(2.3845^n)$. The fourth section presents the conclusions and discusses open problems that may be addressed in future research.

2 Preliminaries

In this section, the main theorems and the model necessary for the algorithm to work will be described.

Model. Our algorithms work in the commonly used QRAM (quantum random access memory) model of computation (Giovannetti, Lloyd and Maccone, 2008), which assumes quantum memory that can be accessed in a superposition.

Tools. We will use the following results in our algorithms:

Theorem 1. *[Quantum Minimum Finding (1996)] Let $a_1,\ldots,a_n$ be integers, accessed by a procedure $\mathcal{P}$. There exists a quantum algorithm that finds $\min_{i=1}^{n}\{a_i\}$ with success probability at least 2/3 using $\mathcal{O}(\sqrt{n})$ applications of $\mathcal{P}$.*

Theorem 2. *[Minimal dominating sets listing (2008)] For any graph G on n vertices, all its minimal dominating sets in X can be listed in time $\mathcal{O}(1.7159^n)$.*

Theorem 3. *[Classical domatic number finding (2008)] There is a classical algorithm that solves* DOMATIC NUMBER *for any graph $G(V,E)$ on n vertices in time $\mathcal{O}(2.8718^n)$ and lists all minimal dominating sets contained in a given $X \subseteq V$ in time $\mathcal{O}(\lambda^{n+\alpha_4|X|})$ with the following values for the weights: $\lambda = 1.148698$ and $\alpha_4 = 2.924811$.*

3 The algorithm

In this section, two quantum algorithms that solve the problem will be described. The simple algorithm solves the problem in time $\mathcal{O}(2.4143^n)$, while the improved algorithm solves it in time $\mathcal{O}(2.3845^n)$.

3.1 Simple algorithm

Our strategy consists of two parts. Firstly, recursively finding all minimal dominating sets in the graph G. The naive deterministic algorithm for listing minimal dominating sets runs in time 2^n, up to polynomial factors. Secondly, we use *Quantum Maximum Finding* to find the largest partition into sets (Ahuja, Kapoor, 1999). This algorithm is based on quantum Grover's search algorithm (Grover, 1996). The analysis of the domatic number finding algorithm is based on (Ambainis, Balodis, Iraids, Kokainis, Prūsis and Vihrovs, 2018), Theorem 1 and is similar to Lawler's classical algorithm for computing the chromatic number (Lawler, 1976) (Fomin, Grandoni, 2005).

Theorem 4. *There is a bounded-error quantum algorithm that solves* DOMATIC NUMBER *in time $\mathcal{O}(2.4143^n)$.*

Proof. The algorithm calculates $d(G)$ for the graph $G(V,E)$ by iterating through all possible subsets $X \subseteq V$. The algorithm seeks to find the largest partition into disjoint dominating sets. The corresponding recursive formula is:

$$d(X) = \max\{d(X \setminus D) + 1 \mid D \subseteq X,\ D \text{ is a minimal dominating set in } G\} \quad (1)$$

Note that for $X \in \emptyset$, $d(X) = 0$.

By replacing classical maximal value search with the quantum maximum finding algorithm in $d(G)$, we obtain a quantum speedup for this problem. Quantum maximum search achieves a quadratic speedup over classical exhaustive search. For each $X \subseteq V$, the quantum algorithm will find $\max_{i=1}^{|X|}\{D_i\}$ in time $\mathcal{O}^*(\sqrt{|X|})$ [1] (Durr and Høyer,

[1] The $\mathcal{O}^*(f(n))$ notation hides a polynomial factor in n

1996). Until we have to iterate through all subsets in G, the running time of the algorithm is bounded by:

$$\sum_{i=0}^{n} \binom{n}{i} i^{\mathcal{O}(1)} \sqrt{2^i} \leq n^{\mathcal{O}(1)} \sum_{i=0}^{n} \binom{n}{i} \sqrt{2^i} = n^{\mathcal{O}(1)} (1+\sqrt{2})^n \in \mathcal{O}(2.4143^n) \square$$

3.2 Improved algorithm

The main idea of the improved algorithm is to precompute all minimal dominating sets (MDS) in $G(V, E)$, where $|V| = n$. The analysis of this algorithm is based on (Ambainis, Balodis, Iraids, Kokainis, Prūsis and Vihrovs, 2018), Theorems 1, 3, and has similarities with Lawler's classical algorithm for computing the chromatic number (Lawler, 1976).

Theorem 5. *There is a bounded-error quantum algorithm that solves* DOMATIC NUMBER *in time* $\mathcal{O}(2.3845^n)$.

Proof. We use the same recurrence as in the proof for Theorem 4 to find the domatic number $d(G)$ for graph $G(V, E)$:

$$d(X) = \max \{d(X \setminus D) + 1 \mid D \subseteq X,\, D \text{ is a minimal dominating set in } G\}$$

Note that for $X \in \emptyset$, $d(X) = 0$. We preprocess the data to enable quick access to all minimal dominating sets. Namely, we create a data structure that can answer two types of queries:

- Given a subset of vertices $X \subseteq V$, what is the number of minimal dominating sets contained in X?
- Given X and i, what is the i^{th} minimal dominating set contained in X (in some fixed ordering)?

Lemma 1. *There is a data structure that can be created in time* $\mathcal{O}^*(2^n)$ *and, given this data structure, the queries of the two types can be answered in time* $\mathcal{O}(n^c)$.

If this data structure has been created, we can find the value of $d(X)$ from equation (1) in time $\mathcal{O}^*(\sqrt{D(X)})$ (where $D(X)$ denotes the number of minimal dominating sets contained in X) using quantum maximum finding from Theorem 1.

We then use this to compute $d(X)$ for all X, in the order of increasing $|X|$. Since $D(X) = \mathcal{O}(\lambda^{n+\alpha_4 i})$, the running time for performing this is of the order at most

$$n^{\mathcal{O}(1)} \sum_{i=0}^{n} \binom{n}{i} \sqrt{\lambda^{n+\alpha_4 i}} = n^{\mathcal{O}(1)} \sum_{i=0}^{n} \binom{n}{i} \lambda^{\frac{n}{2}+\frac{\alpha_4 i}{2}} = \lambda^{\frac{n}{2}} (1+\lambda^{\frac{\alpha_4}{2}})^n n^{\mathcal{O}(1)}$$

$$1.148698^{\frac{n}{2}} (1+1.148698^{\frac{2.924811}{2}})^n n^{\mathcal{O}(1)} \in \mathcal{O}(2.3845^n)$$

It remains to prove Lemma 1. To store all the sets in memory, we will utilise two Hasse diagrams denoted as h_1 and h_2. (A Hasse diagram is a structure with entries for

subsets $X \in V$ in which the entry for X has pointers to the entries for $X - \{u\}, u \in X$.) Each element in h_1 will have the data type 'Node', while each element in h_2 will have the data type 'DNode'.

struct Node
Set: **Set** of **integer**
Subsets: **Set** of Node
Mds: **bool**
MinDomSets: **DNode**
InDNode: **List** of (A**: set,** B**: set,** Ref**: DNode)**

For an element v, v.*Set* contains the set X. v.*Subsets* contains links to 'Node' structures for all subsets of X whose cardinality is one less than the one of X. v.*Mds* is **true** if X is a minimal dominating set and **false** otherwise. v.*MinDomSets* and v.*InDNode* provide links to the entries of Hasse diagram h_2 and are described later, after we describe h_2.

The Hasse diagram h_2 has elements corresponding to pairs of sets A, B with $A, B \subseteq V$ and $\max(i \in A) < \min(j \in B)$. (That is, every element of A must be smaller than every element of B. The corresponding entry describes the number of minimal dominating sets X with $A \subseteq X \subseteq A \cup B$ and provides a way to index them. The elements of h_2 will have the data type 'DNode'.

struct DNode
A: **Set** of **integer**
B: **Set** of **integer**
Count: **integer**
Mds: **bool**
Subsets: **Set** of **DNode**

We say that for each element v_1, v_2 in h_2: $v_2 \prec v_1$ iff

$$\exists b \in v_1.B : (v_2.A = v_1.A \cup b) \wedge (v_2.B = v_1.B \setminus \{x \in v_1.B | x \leq b\})$$

In h_2, the fields have the following content. A, B represent vertex sets in v_1, while *Subsets* contains links to all v_2 with $v_2 \prec v_1$. We mark *Mds* as True iff A is minimal dominating set. *Count* contains the number of X with $v_1.A \subseteq X \subseteq v1.A \cup v1.B$.

We note that if X is such that $v_1.A \subset X \subseteq v1.A \cup v1.B$, then X satisfies $v_2.A \subseteq X \subseteq v2.A \cup v2.B$ for exactly one v_2 with $v_2 \prec v_1$. Namely, this will be the node v_2 with $v_2.A = v_1.A \cup \{c\}$ where c is the smallest element of $v_1.B$ and $v_2.B = \{x | x \in v_1.B \wedge x > c\}$. Thus, the set of X with $v_1.A \subset X \subseteq v1.A \cup v1.B$ is a disjoint union of the sets of X with $v_2.A \subseteq X \subseteq v2.A \cup v2.B$ for all $v_2 \prec v_1$. All those v_2 are enumerated by the *Subsets* field.

Lastly, we describe the references from h_1 to h_2. There are two types of them.

The first type, denoted as *MinDomSets*, finds the element of type **DNode** that refers to i-th MDS within h_2. **Node**.*MinDomSet* = **DNode** iff **Node**.*Set* = **DNode**.B and **DNode**.$A = \emptyset$.

The second type of reference, called *Ref*, is stored in the *IsDNode* list. For every N of **Node** objects, the *IsDNode* list maintains pairs of sets denoted as $(\mathcal{A}, \mathcal{B})$. These pairs satisfy the condition that $\mathcal{A} \cup \mathcal{B} = N.Set$. Furthermore, it is required that an object of type **DNode** has been previously instantiated within h_2. This prevents the occurrence of duplicates in h_2.

Next we describe Algorithm 1 that find i-th minimal dominating set in polynomial time. The algorithm 1 consists of two parts.

In the first part, we identify the subset S in the Hasse diagram h_1 in time $\mathcal{O}(n^2)$ using function FINDSUBSET.

In the second part, the algorithm refers to **DNode** object in h_2 using **Node** *MinDomSet*. After this, the search for the i-th subset in h_2 starts. We check all *Subsets* of the current **DNode** element until the sum of all previous *Count* values is less than i. When this sum equals or exceeds i, the algorithm descends one level lower. This process continues until **DNode** *Mds* is not *True*. When navigating h_2, this search is confined to a maximum of n *Subsets*, at most n times, for each element in B. Therefore the second part operates in time $\mathcal{O}(n^2)$.

Algorithm 1 Finding the i-th minimal dominating set among subsets of S

Input: S - subset of V, i - index for the minimal dominating set

```
1: function FindSubsetMinDomSetByIndex(S: set, i: integer)
2:     S ← FindSubset(S).MinDomSets                          ▷ S: DNode
3:     while S.Mds = False do                                ▷ Search in h2
4:         c ← 0
5:         for subset in S.Subsets do                        ▷ subset: DNode
6:             if c+subset.Count ≥ i then
7:                 S ← subset
8:                 i -= c
9:                 c ← 0
10:                break
11:            else
12:                c ← c + Count
13:    return S.A
```

For the purpose of precomputation all minimal dominating sets for G, we execute Algorithm 2. This precomputation algorithm is designed to store all subsets of V within the h_1. The naive algorithm for finding the i-th minimal dominating set of S in h_1 requires exponential time to visit all subsets of S. Consequently, we introduce h_2 and the procedure D to calculate the count of minimal dominating sets that include the elements of set A and potentially some elements from set B:

$$\mathrm{D}(A,B) = \begin{cases} \sum_{i=0}^{k<n} \mathrm{D}(A \cup b_i, \{b_{i+1}, \ldots, b_k\}) & \text{, if } A \text{ is not MDS } \textbf{and } B \neq \emptyset \\ 1 & \text{, if } A \text{ is MDS} \\ 0 & \text{, otherwise.} \end{cases}$$

$$B = \{b_0, \ldots, b_k\} = \{b \mid \forall i, j \geq 0 (i < j \rightarrow b_i < b_j)\}$$

To obtain all MDS for subset S we run D($\emptyset, S$).

Algorithm 2 Precomputation of minimal dominating sets

Input: $G(V, E)$ - input graph.
Output: h_1, h_2 with precomputed MDS.

1. Generate all subsets of V and place them within the h_1.
2. Execute the algorithm for generating minimal dominating sets from Theorem 2. When a certain set D is returned:
   2.1. Find D in h_1, using Function FINDSUBSET.
   2.2. Mark *Mds* as True in found D node in h_1.
3. Execute Procedure PRECOMPUTEMINDOMSETS.

The procedure PRECOMPUTEMINDOMSETS is implemented in Algorithm 3, which creates h_2 using h_1. For each subset $S \subseteq V$, the algorithm executes D($\emptyset, S$) and stores each term of D($\emptyset, S$) in h_2. To find all subsets of V we run Function GETSUBSETS. While $\forall a(a \in A \rightarrow \forall b(b \in B \rightarrow a < b))$ holds true, we state that $A \subseteq \{0, \ldots, m-1\}$ and $B \subseteq \{m, \ldots, n-1\}$. Consequently, there are 2^m ways to choose A and 2^{n-m+1} ways to choose B, with n ways to choose the value of m. Therefore, Algorithm 3 runs in time $\mathcal{O}^*(n \cdot 2^m 2^{n-m+1}) = \mathcal{O}^*(2^n)$. Now, we just need to demonstrate that the auxiliary functions for the Algorithm 3 will operate in polynomial time.

Note that the Algorithm 3 requires exponential space $\mathcal{O}^*(2^n)$.

Algorithm 3 Implementation of the PRECOMPUTEMINDOMSETS Procedure

Input: P: Node - pointer on h_1 head element

```
1: procedure PRECOMPUTEMINDOMSETS                          ▷ Creates h2
2:     for subset in GETSUBSETS(P.Set) do                  ▷ Subsets in h1
3:         A ← ∅
4:         B ← subset.Set = {x | ∀i, j ≥ 0(i < j → x_i < x_j)}
5:         N ← new DNode(
6:                     A, B, Count : 0, Mds : ISMDS(subset), Subsets : ∅)
7:         subset.MinDomSets ← N
8:         INSERTDNODEREF(A, B)
9:         D(A, B)
```

Prior to adding an element to the h_2, Algorithm 3 performs a verification step by invoking the FINDDNODE function. This function checks for the presence of the current element within h_2. Unless each $a \in \mathcal{A}$ and $b \in \mathcal{B}$, $a < b$, the size of the *IsDNode* list will be at most $\mathcal{O}(n)$. Therefore FINDDNODE operates in polynomial time. If the element is found, the *Count* attribute of the corresponding **DNode** instance is incremented by the value from the current subset. In cases where the element is not found, the algorithm adds reference *Ref* from h_1 to h_2 using INSERTDNODEREF function. After that it computes the *Count* value via a recursive formula for the D function.

```
10: procedure D(A, B = {x | ∀i, j ≥ 0(i < j → x_i < x_j)})
11:     if IsMds(A) then
12:         return 1
13:
14:     for k = 0; k < |B|; k++ do
15:         A ← A ∪ {x_k}
16:         B ← {x_{k+1}, x_{k+2}, ..., x_n}
17:         R ← FindDNode(A, B)                          ▷ Type: DNode
18:
19:         if R = ∅ then
20:             N.Subsets add new DNode(
21:                 A, B, Count : 0, Mds : False, Subsets : ∅)
22:
23:             InsertDNodeRef(A, B)
24:             N.Count += D(A, B)
25:         else                                         ▷ Computed early
26:             N.Subsets add R
27:             N.Count += R.Count
28:     return 0
29:
30: function GetSubsets(node: Node)
31:     if node = ∅ then
32:         return
33:     yield node                                       ▷ Type: Node
34:     for subset in node.Subsets do
35:         GetSubsets(subset)
36:
37: function InsertDNodeRef(A: Set, B: Set)
38:     S ← FindSubset(A ∪ B)                            ▷ Type: Node
39:     S.InDNode add (A, B)
40:
41: function IsMds(S: Set)
42:     S ← FindSubset(S)                                ▷ Type: Node
43:     if S = ∅ then                                    ▷ No such set in diagram
44:         return False
45:     return S.Mds
46:
47: function FindDNode(A: set, B: set)
48:     S ← FindSubset(A ∪ B)                            ▷ Type: Node
49:     for (A, B, Ref) in S.InDNode do
50:         if A = A and B = B then
51:             return Ref                               ▷ Type: DNode
52:     return ∅
```

Before calculating each term of the recursive function D for $(A, B) = (\emptyset, S)$, the Algorithm 3 checks whether the set A is MDS using the IsMds function. If the current set A is MDS, then the algorithm will return 1. Otherwise, it will call the function D until $B \neq \emptyset$.

Next we describe how to implement FindSubset, using Algorithm 4. Note that each set contains no more than n edges from current node to it subsets. When navigating the Hasse diagram h_1, our search is confined to a maximum of n sets. Therefore Algorithm 4 runs in time $\mathcal{O}(n^2)$.

Algorithm 4 Finding subsets in Hasse diagram h_1

Input: D - subset, G - pointer on h_1 head element

```
1: function FindSubset(D: set)
2:     S ← G
3:     while S.Set ≠ D do
4:         if S.Subsets = ∅ then                ▷ No such set in diagram
5:             return ∅
6:         for s ∈ S.Subsets do
7:             if D ⊆ s then
8:                 S ← s
9:                 break
10:    return S
```

4 Conclusion and open problems

The Domatic Number problem holds significant importance in graph theory and finds practical applications in network optimization. The quantum algorithm for this problem recursively searches for minimal dominating sets within a graph. At each recursion level, sets are excluded from the current vertex set until all vertices are covered. The algorithm determines the maximum recursion level where each excluded set remains dominating.

Our findings revel that the one of the best-known classical algorithm solves the problem with a time complexity of $\mathcal{O}(2.7139^n)$ (van Rooij, 2010) using a dynamic programming approach. Quantum maximum search provides a quadratic speedup for enumerating all minimal dominating sets, thereby achieving a faster solution to the Domatic Number problem.

Our newly presented quantum algorithms demonstrate the power of quantum computing to accelerate classical algorithms and achieve improved time complexity with relatively straightforward implementation. Combining dynamic programming with quantum algorithms improves the evaluation of the Domatic Number algorithm.

A key achievement in this work includes Theorem 4 and 5. Employing the Quantum Maximum Finding algorithm, we have demonstrated algorithms that solve the Domatic Number problem in time $\mathcal{O}(2.4143^n)$, which we have further improved to

$\mathcal{O}(2.3845^n)$. Gaspers and Li (2023) have independently developed a quantum algorithm with the complexity of $\mathcal{O}((2-\epsilon)^n)$.

It would be interesting to explore whether there exists an algorithm that can further improve this complexity. However, limitations include the requirement for exponential space, which may not be practical for large-scale graphs. This presents a crucial area for further research.

Acknowledgment. AA was supported by QuantERA ERANET Cofund project QOPT (Quantum algorithms for optimization).

References

Ahuja, A., Kapoor, S. (1999) A Quantum Algorithm for finding the Maximum *Quantum Physics* at `arxiv.org/abs/quant-ph/9911082`

Ambainis, A., Balodis, K., Iraids, J., Kokainis, M., Prūsis, K., Vihrovs, J. (2018). Quantum Speedups for Exponential-Time Dynamic Programming Algorithms, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1783-1793. doi:10.1137/1.9781611975482.107

Chang., G.J. (1994) The domatic number problem, *Elsevier*, 125, issues 1-3. doi: 10.1016/0012-365X(94)90151-1

Durr, C., Høyer, P. (1996). A Quantum Algorithm for Finding the Minimum, preprint, available at `arXiv:quant-ph/9607014`

Fomin, F.V., Grandoni, F., Pyatkin, A.V., Stepanov, A.A. (2005) Bounding the number of minimal dominating sets: A measure and conquer approach. *Algorithms and Computation* 3827. doi:10.1007/11602613_58

Fomin, F.V., Grandoni, F., Pyatkin, A.V., Stepanov, A.A. (2008). Combinatorial bounds via measure and conquer: Bounding minimal dominating sets and applications. *ACM Transactions on Algorithms* **5**(1), 9:1-9:17. doi:10.1145/1435375.1435384

Fomin, F.V., Kratsch, D. (2010). Exact Exponential Algorithms. *Springer* p.36. doi:10.1145/2428556.2428575

Gaspers, S., Li, J.Z. (2023). Quantum Algorithms for Graph Coloring and other Partitioning, Covering, and Packing Problems. preprint, available at `arXiv:2311.08042`

Giovannetti, V., Lloyd, S., Maccone, L. (2008). Quantum Random Access Memory. *Phys. Rev. Lett.* **100**, p. 160501. doi: 10.1103/PhysRevLett.100.160501

Grover, L. (1996). A Fast Quantum Mechanical Algorithm for Database Search. *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing.* p.2112-229. doi: 10.1145/237814.237866

Jiguo, Y., Qingbo, Z., Dongxiao, Y., Congcong. C., Guanghui, W. (2014). Domatic partition in homogeneous wireless sensor networks. *Elsevier* doi: 10.1016/j.jnca.2013.02.025

Lawler, E.L. (1976). A note on the complexity of the chromatic number problem. *Information Processing Lett.* **5**, 66–67, doi: 10.1016/0020-0190(76)90065-X

Repko, I. (2023). Quantum Algorithms for Domatic Number Finding Problem. preprint, available at `https://dspace.lu.lv/dspace/handle/7/63279`, last viewed <14.02.2024>

Riege, T., Rothe, J. (2005) An Exact 2.9416ˆn Algorithm for the Three Domatic Number Problem. *Mathematical Foundations of Computer Science* 3618. doi: 10.1007/11549345_63

Riege, T., Rothe, J., Spakowski, H., Yamamoto, M. (2007) An improved exact algorithm for the domatic number problem *Information Processing Letters* 101, p.101-106. doi: 10.1016/j.ipl.2006.08.010

Van Rooij, J.M.M. (2010). Polynomial Space Algorithms for Counting Dominating Sets and the Domatic Number. *Algorithms and Complexity, Lecture Notes in Computer Science* **6078**, pp. 73-84. doi: 10.1007/978-3-642-13073-1_8

Received February 14, 2024 , revised June 16, 2024, accepted August 16,2024

Baltic J. Modern Computing, Vol. 12 (2024), No.3, 270-285
https://doi.org/10.22364/bjmc.2024.12.3.04

Mathematical and Computer Simulation of the Process of Movement of Respirable Dust Particles in the Working Area

Yevhenii LASHKO[1], Olha CHENCHEVA[1], Ivan LAKTIONOV[2], Dmytro RIEZNIK[1], Nadiia HALCHENKO[1]

[1]Institute of Education and Science in Mechanical Engineering, Transport and Natural Sciences, Department of Civil and Labour Safety, Geodesy and Land Management, Kremenchuk Mykhailo Ostrohradskyi National University, str. Universytetska, 20, Kremenchuk, 39600, Ukraine
[2]Faculty of Information Technologies, Department of Computer Systems Software, Dnipro University of Technology, 19 Dmytra Yavornytskoho av., Dnipro, UA49005, Ukraine

evgeny.lashko.lj@gmail.com, chenchevaolga@gmail.com, Laktionov.I.S@nmu.one, 2411dimareznik@gmail.com, nadingal9@gmail.com,

ORCID 0000-0001-9691-4648, ORCID: 0000-0003-2659-177X ORCID 0000-0002-5691-7884, ORCID 0000-0001-7857-6382, ORCID 0000-0003-1258-6136

Abstract. Most air purification systems are formed on the basis of the "modular principle" using a waste-free production scheme, standardized dust collection equipment and ventilation systems. The disadvantages of such a complex, which is assembled from heterogeneous purification equipment, are the large overall dimensions of the devices, low individual performance and low gas flow rates in the devices, which limit the ability to purify large volumes of air, and therefore the task of preliminary verification of their capabilities arises. Modern computer simulation software makes it possible to study the movement of microscopic particles and determine the stable patterns of this process. This study focuses on the mathematical and computer simulation of the process of movement of respirable dust particles in the working area, based on the principles of designing efficient modular devices that maximize the use of centrifugal force to improve the performance of dust and gas cleaning equipment. The object of study is the proposed air purification device as a separate part of the overall purification complex. The subject of the study is computer simulation of the movement of dust particles in the air flow. The scientific and practical value of the research results is that for the first time the regularities of aerodynamic processes occurring in the cylindrical body of the "centrifugation module" were determined, which were obtained by mathematical and computer simulation methods, which confirms the effectiveness of air purification by the proposed device and makes it possible to introduce it into mass production.

Keywords: simulation, air purification, technological equipment, dust particles.

1. Introduction

Many production processes and air purification facilities are based on the "modular principle", which uses waste-free production schemes, standardized dust collection equipment and ventilation systems. The disadvantages of such a "modular complex", assembled from heterogeneous purification equipment, are the large overall dimensions of the devices, their low individual performance and low gas flow rates in the devices, which limit their capabilities when it comes to purifying large volumes of air. Therefore, the task of preliminary verification of the capabilities of any proposed purification systems arises. Modern methods of mathematical simulation and computer experimentation software allow us to study the movement of microscopic particles and determine the stable patterns that arise, as well as analyze not only the parameters of individual particles but also the macroscopic parameters of aerodynamic systems in general.

The main part of this complex of gas purification equipment is the process of dry particle separation in cyclones. The centrifugal force acting on the particle determines the equilibrium position of the particle in the flow and its separation (Birkhoff, 2015):

$$F = c_1 d_p^3 \left(\rho_p - \rho_{sp}\right) \omega^2 / r, \tag{1}$$

where c_1 is a constant; d_p is the diameter of the particle; ρ_p and ρ_{sp} are the densities of the particle and the space; ω is the angular velocity of the particle; r is the radius of its rotation.

Formula (1) shows, in particular, that high centrifugal forces and, consequently, high efficiency of the separation process can be achieved at high angular velocity of particles. However, it is worth noting that increasing the rotational speed in cyclones does not produce the desired effect due to the fact that the kinetic energy of turbulence increases, which intensifies the process of reverse mixing of the separated dust stream with the "clean" gas leaving the apparatus. In addition, at high inlet velocities, most of the dust falls out at a relatively short distance from the inlet, where it accumulates in large quantities on the cylindrical wall of the cyclone and increases its resistance. Therefore, the operating value of the cyclone inlet velocity is limited depending on the cyclone diameter.

Another important aspect is the origin and nature of the particles themselves. For example, machining of carbon-containing composite materials is characterized by a significant dust emission, which contains carbon fiber residues, nanotubes, coal dust and epoxy residues. At the same time, the processing of this type of material with an abrasive tool increases the amount of dust that is carried away from the cutting zone (Bayraktar et al., 2016). Such phenomena lead to the occurrence of occupational diseases, especially in the absence of personal protective equipment for workers in contact with carbon materials.

Existing computational complexes allow simulation the behavior of particles in various spatial configurations of the study area (Bai, 2017), while calculating statistical estimates of macroscopic parameters (density, temperature, pressure) in elementary volumes (Kumar et al., 2020; Biliaieva et al., 2019).

However, all of these studies have some, in our opinion, significant limitations that do not allow us to speak about the convergence of computer simulation results with the data obtained during the field experiment, since a number of important factors were not taken into account when entering the initial data, and most quantitative indicators need to be clarified. These factors include the volumetric dust consumption and its fractionation, which are the basis for the calculation as the initial data for this research.

This paper deals with the mathematical and computer simulation of the process of movement of inhaled dust particles in a purification device, based on the principles of designing efficient modular devices that use centrifugal force to increase the performance of dust and gas cleaning equipment. The object of study is the proposed three-dimensional model of the device ("module") for air purification, which is a component of the overall purification complex. The subject of the study is the simulation of the movement of dust particles in the air flow of the specified device. The main tasks that arise when designing such "modules" are to increase the rotation speed and residence time of the contaminated flow inside such a device. Therefore, this work focuses on mathematical and computer simulation, which has become the main tool for studying complex processes and systems today, and on which modern approaches to the design of products for various purposes are based. The article contains a mathematical description of the principles of creating air purification systems, designing a 3D model of a purification device, and computer simulation of aerodynamic processes occurring in its working area.

The scientific and practical value of the research results of this article lies in the fact that for the first time the regularities of aerodynamic processes occurring in the cylindrical body of the "centrifugation module" were determined by mathematical and computer simulation methods, which confirms the effectiveness of air purification by the proposed device and makes it possible to introduce it into mass production.

2. Methods, tools and approaches to research

The main research results of the article are based on the methods of critical analysis and logical generalization of the known results of scientific research in the field of simulation the technical mechanics of liquids and gases (Liu et al., 2019; Xiu et al., 2020; Zhou et al., 2022).

The research of this article is a logical continuation of the author's own theoretical and experimental studies in the field of mathematical and computer simulation, which are reflected in scientific articles (Chencheva et al., 2023; Salenko et al., 2020).

The results obtained were validated by conducting a full-scale experiment in the working area.

Computer simulation of particle motion can be divided into three stages:

1) determining the initial conditions;

2) changing the position of particles in space;

3) visualization of the results.

The initial conditions are the initial position of the particles, the number of particles, the initial vectors of the directions of movement in space, and the boundaries of the zones (in the form of continuous functions). The calculation step (the distance the particle travels in a given time period) and the calculation accuracy used to solve the problem of particle reflection from the zone boundary are also determined.

Each particle moves towards the boundary of the zone. Determining the new position of the particle after reflection involves the following steps:

1) selecting the boundary of the zone that the particle crosses in its trajectory;

2) determining the point of intersection of the particle trajectory and the boundary;

3) selecting the shortest distance from the initial position of the particle to the intersection point;

4) calculating the normal for the function at the intersection point;

5) calculation of the new direction vector.

All these parameters are calculated after each step (iteration).

Today, similar tasks are successfully solved in rolled steel cyclone chambers, where dispersed materials are subjected to heat treatment. The flow velocity at the chamber inlet reaches 190 m/s at a Reynolds number of Re=100000. Significantly higher centrifugal accelerations are achieved in centrifuges, machines that separate heterogeneous systems in a high-intensity centrifugal field, which exceeds the acceleration of natural dust particle deposition by tens of thousands of times. For example, in gas centrifuges for isotope separation, due to the high rotor speed, the linear velocity at the periphery can exceed 600 m/s. In this case, the product is concentrated near the chamber wall under high pressure, and a so-called vacuum core is formed in the rotor axis, which provides additional axial gas circulation inside the rotor.

However, these devices are very complex and expensive. For example, the main working element in a centrifuge is a hollow rotor that rotates rapidly around an axis, with finishing coatings, heat treatment and precision manufactured to aviation standards. The high quality of workmanship and cost of a whole range of auxiliary devices, such as rotor supports, magnetic bearings, motors, etc. Each of these devices is unique. Consistent operation of centrifuge components requires fine-tuning and highly skilled maintenance. These features practically exclude the use of centrifuges for the separation of suspensions from flue and corrosive gases, however, the method of achieving powerful centrifugal fields created by them remains a promising task for research and development.

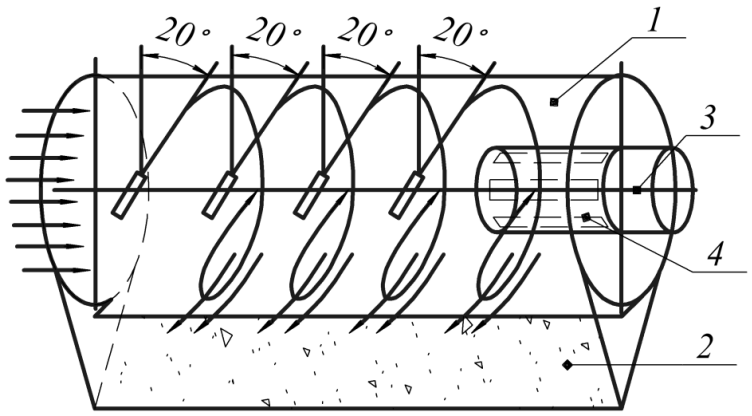


Figure 1. Schematic of suspended solids separation by gas centrifugation in a fixed vessel using high-speed hydrodynamic jets

In the proposed "centrifugation module", in order to preserve the above positive qualities of the centrifuge, it was decided to use a fixed body 1 – a cylinder with a cut-out 90° sector, which at the same time serves as a negative electrode. The proposed device is equipped with a rotor of the drive mechanism 3 with a positive electrode 4 placed on it. The cylindrical body 1 is mounted on the side sheets of the removable hopper 3 at a level 20° below the horizontal axis of the fixed body 1. On the right side, the side sheet is attached flush, and on the left side, it is overlapped with the body. In this case, on the left

side, a part of the fixed body 1 in the form of a wing guided by the flow stream hangs over the collection hopper 2 (Fig. 1). During the centrifugation of gas flows, the "wing" directs the sludge and dripping liquid deposited on the walls of the fixed body 1 into the removable hopper 2.

The process of centrifugation of flue and corrosive gases in the proposed device is carried out by injecting hydrodynamic high-speed jets directed tangentially inside the fixed casing and co-directed with the flows of gases to be cleaned. To ensure the linear movement of the gas flow and the formation of a "vacuum core" along the casing axis, the hydrodynamic jets are fed with a 20° inclination to the vertical plane in the direction of the main flows inside the casing. The spiral motion of the jet allows the compacted sludge to be continuously discharged from the internal cavity of the fixed casing into the removable hopper.

The regularities of hydrodynamic processes occurring in the cylindrical body of the centrifugation module can be considered by analogy with cyclone piercing chambers operating at high flow rates.

All the mathematical dependencies used for the calculation are well-known in the context of applied technical mechanics of liquids and gases (Tannehill et al., 1997). At the same time, the equations are adapted in accordance with the research goal, taking into account the geometric parameters of the designed device and the conditions of its full functioning.

The volume of gas flow passing through the cross-section of the enclosure can be determined from the equation:

$$V_g = 2\pi r_o w_{mo} L = 2\pi r_c w_{mk} L_t, \tag{2}$$

where r_0 is the body radius; L_t is the length of the active turn of the gas vortex; r_c is the current radius; $w_{mk}=w_{mo}r_o/r_c$ is the radial flow velocity.

The resulting velocity vector is shifted relative to the angular velocity vector by an angle that causes the gas flow to move towards the axis and dust particles and condensate droplets to move towards the periphery of the chamber:

$$tg\alpha = w_{mk} / w_{nk} = w_{mo} / w_{no}, \tag{3}$$

where $w_{mk}=w_{mo}r_o/r_k$ is the angular velocity of the vortex flow.

The conditions for the movement of a particle along a circular trajectory are described by the equation that reflects the equality of the gas pressure force on the particle and the centrifugal force acting on it:

$$\psi \frac{\pi d_p^2}{4} \frac{w_{mk}^2}{2} \rho_g = \frac{\pi d_p^3}{6} \frac{w_{mk}^2}{r_c} \rho_p, \tag{4}$$

where $\psi=24v_p/d_p w_{mk}\rho_g$, and the physical properties of the gas are assumed to be at an average temperature.

The head loss during the flow into the apparatus is estimated by the formula:

$$\Delta P_1 = \frac{\rho_g}{2}\left(\frac{V_g}{\mu F_0}\right)^2, \tag{5}$$

where μ=0.85 is the flow coefficient; F_0 is the smallest cross-section of the Laval nozzle.

The head loss during vortex formation can be calculated using the formula:

$$\Delta P_2 = \frac{\rho_g}{2}\left[\left(\frac{D_0}{d_1}\rho\right)^2 - 1\right] w_{no}^2, \tag{6}$$

where D_0 is the diameter of the cylindrical body; d_1 is the diameter of the separator outlet.

The main hydrodynamic drag is concentrated in the fixed chamber, so a simplified formula can be adopted for the region Re=100000 (at an air velocity at the chamber inlet in the range of 190 m/s).

$$\Sigma\Delta P = 0.07\left(\frac{\Sigma F_0}{F_k}\right), \tag{7}$$

where ΣF_0 is the total area of the openings for the outlet of high-speed compressed air jets; F_K is the cross-sectional area of the chamber.

The leading role in the centrifugation process in the analyzed "module" is played by hydrodynamic accelerators with Laval nozzles, in which the compressed air outflow rate exceeds the Mach number (M>1). The opening angle of the jet plume in the open atmosphere is 23°, and in the compressed conditions of the plant vessel it reaches 50°. The compressed air flow rate from the Laval nozzle of a hydrodynamic accelerator can be determined by the formula for adiabatic flow:

$$\omega_1 = \varphi\sqrt{2g\cdot\frac{k}{k-1}\cdot\frac{P_1}{\gamma_g}\left[1-\left(\frac{P_2}{P_1}\right)^{\frac{k-1}{k}}\right]}, \tag{8}$$

where γ_g is the specific gravity of the gas in front of the nozzle at pressure P_1 ; φ is the leakage coefficient (for a nozzle with a cylindrical part and an angle of β=45° at l/d=0.18 φ=0.75, at l/d=0.56 φ = 0.9); k is the adiabatic coefficient (for two-atom gases and air, k=1.4); g is the acceleration of gravity; P_1 is the compressed air pressure before the nozzle; P_2 is the pressure before the nozzle outlet, equal to 101300 Pa.

Let us determine the air outflow rate from the nozzle of a hydrodynamic accelerator designed as a Laval nozzle and the compressed air flow rate per second under the conditions of air outflow into an environment with a pressure close to atmospheric pressure, i.e., where the pressure is lower than the critical pressure. In this mode of leakage, the pressure at the outlet of the nozzle is set equal to the critical pressure, and the leakage rate is equal to the critical rate, and the flow rate is maximum. The critical flow rate can be determined by the formula:

$$\omega_{cr} = \varphi\sqrt{2\frac{k}{k+1}RT_0}, \tag{9}$$

where R is the gas constant; T_0 is the gas temperature.

For comparison, the sound velocity at the nozzle outlet:

$$a_s = \varphi\sqrt{KRT_2}, \tag{10}$$

where , $T_2 = T_0\beta_{cr}^{\frac{k-1}{k}}$ $\beta = \frac{P_{av}}{P_0}$ are pressure ratios.

The compressed air consumption per unit hydrodynamic accelerator is determined by the formula:

$$m = m_{\max} = f\sqrt{2\frac{k}{k+1}\left(\frac{2}{k+1}\right)^{\frac{2}{k+1}}\cdot\frac{P_1}{V_0}}, \tag{11}$$

where P_1 is the air pressure in front of the nozzle; f *is* the nozzle cross-section; V_o is the specific volume of air in front of the nozzle.

$$V_0 = \frac{h_1 - u_1}{P_1}, \tag{12}$$

Here h_1 is enthalpy; u_1 is internal energy.

For two-atom gases, we have the following parameters under standard conditions: enthalpy h_1=283.2 kJ/kg; internal energy u_1=209.2 kJ/kg. From the relation for the ideal state of the gas $h=u+RT$, we find the gas constant:

$$R = \frac{h-u}{T} = \frac{283.2 - 209.1\cdot10^3}{297.6} = 2524 \text{ kJ/kgK}, \tag{13}$$

where T=273.6+20=293.6 K.

The ratio of the leakage pressure P_l= 0.5 MPa and the space into which the leakage occurs P_{av}=0.1 MPa is β=0.5/0.1=5. Then the critical pressure for air P_{cr}=0.528 MPa. Let's check whether the critical velocity at the nozzle outlet is really established:

$$P_0 = P_{cr} - P_l = 0.528 - 0.5 = 0{,}028 \text{ MPa}. \tag{14}$$

Therefore, the pressure of the space is lower than the critical pressure and the jet speed should be close to the sound speed:

$$a_s = \sqrt{KRT_{air}} = \sqrt{1.4 \cdot 252.4 \cdot 293.6} = 322.1 \text{ m/s} \quad (15)$$

.

The specific volume of air at a pressure P_l=0.5 MPa is determined by the formula:

$$V_{air} = \frac{h-u}{P_l} = \frac{(283.2-209.1)\cdot 10^3}{5\cdot 10^5} = 0.148 \text{ m}^3/\text{kg} \quad (16)$$

The specific volume of air inside the accelerator cone, i.e. in the leakage zone, is:

$$V_2 = \frac{h-u}{P_2} = \frac{(283.2-209.1)\cdot 10^3}{1\cdot 10^5} = 0.741 \text{ m}^3/\text{kg} \quad (17)$$

Changes in compressed air temperature:

$$T_2 = T_0 \frac{2}{k+1} = 293{,}6 \frac{2}{1{,}4+1} = 244{,}6 \text{ K} \quad (18)$$

Let's determine the actual air leakage rate using the formula:

$$\omega_{cr} = \sqrt{2\frac{k}{k+1}RT_0} = \omega_2 = \sqrt{2\frac{1{,}4}{1.4+1}252.4\cdot 293.6} = 272.2 \text{ m/s} \quad (19)$$

.

Thus, we can assume that the leakage rate is set equal to the local sound speed $a_{зв}$.

With a nozzle diameter of d=15 mm, the calculated area of the outlet section is f=0.000176 m^2. Then the compressed air consumption per unit hydrodynamic accelerator:

$$m = 0{,}000176\sqrt{2\frac{1{,}4}{1.4+1}\left(\frac{2}{1.4+1}\right)^{\frac{2}{1.4+1}}} \cdot \frac{5\cdot 10^5}{0.0139} = 0.1348 \text{ kg/s or } 485.3 \text{ kg/h} \quad (20)$$

.

After returning to normal conditions, the volume flow rate for the individual hydrodynamic accelerator:

$$V = 483.5 \cdot 1.293 = 627.5 \text{ m/h}^3 \quad (21)$$

Compressor units are selected based on the total flow rate V. The total volumes of the receivers are selected to be 40 % larger than the output capacity of the compressors.

The main input data for the calculation are grouped in Table 1. The components of these data are the geometric parameters of the device and quantitative aerodynamic parameters.

Table 1: Initial data of preprocessing

Calculated area of the original section	0,000176 m^2
Specific air volume inside the accelerator cone	0.741 m/kg^3
Changes in compressed air temperature	244.6 K
Volumetric flow rate	485.3 kg/h
Dust fractionation	1–5 μm

Additionally, the phenomenon of gravity in the corresponding spatial computational grid, as well as dust fractionation, are taken into account. A limitation of the simulation is the representation of dust particles in the form of a sphere. At the same time, taking into account the irregular shape of the particles can only be considered as each individual act of its single interaction, which does not allow achieving the research goal of testing the performance and efficiency of the proposed device.

3. Research results

3.1. Computer model

Thus, the main technical characteristics of the centrifugation unit, which can be used in a "modular complex" for gas purification instead of typical cyclone devices with a larger diameter, have been determined. In particular, the speed at the nozzle outlet of the hydrodynamic device is more than 270 m/s. Further search for a basic unit was carried out with a view to ensuring a higher throughput capacity for gas flows, as well as increasing the efficiency of gas purification while reducing metal consumption, energy and cost costs.

The result of solving the set tasks using mathematical and computer simulation methods is the proposed installation for deep dust and sludge collection, which is assembled according to a modular scheme. The "modular complex" uses a single body – a larger diameter pipe with supply and discharge pipes and hoppers for dust and sludge removal. The casing is divided into separate component sections, such as a dust collecting chamber, a catalytic reduction chamber, a multifunctional adsorption "centrifugation chamber", and a draft blower unit, which is based on a gravity chamber for condensate and sludge collection.

The device comprises a large diameter cylindrical casing divided into 4 component modules (I, I, III, IV). Each module performs a specific function as a gas separation and purification stage and can be completed with a "modular complex" depending on the production needs.

Module I is a dust settling chamber for primary gas cleaning. It consists of an inlet pipe 1 with an elbow inserted at an angle of 30° into the body 2 of the dust removal chamber, a baffle diaphragm 3, a collection hopper 4 and a pneumatic conveying pipe 5. The velocity in the dust collecting chamber is within 10 m/s. Dust particles are deposited due to gravitational forces and changes in the direction of flow.

Thus, in the proposed multi-sectional dust and gas collection module, gases are separated and purified using high-speed compressed air flows at high centrifugal velocities corresponding to the level of centrifugation. Hydrodynamic accelerators installed

tangentially at an angle of 20–25º to the vertical introduce high-speed compressed air flows into the internal cavity of the cylindrical casing and create a vortex (swirling flow), creating a vacuum (thrust) that reduces the overall aerodynamic drag of the plant. The use of dry catalytic neutralization of chemical impurities with a choice of adsorbents extends the versatility of the "modular complex", especially when capturing dioxins and polyaromatic hydrocarbons.

The possibility of dosed dispersion of water or adsorbents allows for the combination of dry, condensation and wet dust collection processes by selecting the optimum adsorbents depending on the composition of flue and corrosive gases.

The last stage of gas purification is a bladeless draft blower. In addition to gas purification and cooling, before being released into the atmosphere, the gas is provided with thrust throughout the cavity of the modular complex, which allows it to operate autonomously without the use of heavy thrusting devices in the high-temperature zone. The autonomous operation of the thrust-blower unit is ensured by the kinetic energy of the vortex flows in the centrifugation chamber and in the bladeless ventilation unit. The energy of the vortex flows is supplied by compressed air from the compressor unit.

Additionally, a device for purification of gas media, including positive and negative electrodes connected to a source of electrical energy, creates an electric field in the interelectrode space, and is characterized by the fact that the positive electrode is made in the form of a flat cylindrical body and is installed with the possibility of moving in the opposite direction of the gas space supply relative to the fixedly installed negative electrode and creating an electric field between the electrodes.

In addition, the device has a number of features that characterize it in certain cases of its execution, specific forms of its material embodiment or special conditions of its use, namely:

– the positive electrode can be mounted on a platform that is fixed to the rotor of the drive mechanism;

– the platform on the inner surface of the positive electrode can be equipped with a power supply, a voltage rectifier, a generator and a voltage multiplier connected in series;

– the device can be equipped with a hopper for collecting particles deposited on the positive electrode.

The technical result achieved by using this set of essential features of the device is that the movement of the positive electrode relative to the fixed negative electrode, which is the device body, provides an increase in the efficiency of polarization of particles of the contaminated gas space and their deposition on the outer surface of the positive electrode, as well as the possibility of continuous cleaning of this surface of the positive electrode without the need to stop the operation of the device.

The model of the proposed device as an initial structural element in section is shown in Fig. 2. It was built using SolidWorks 3D simulation software.

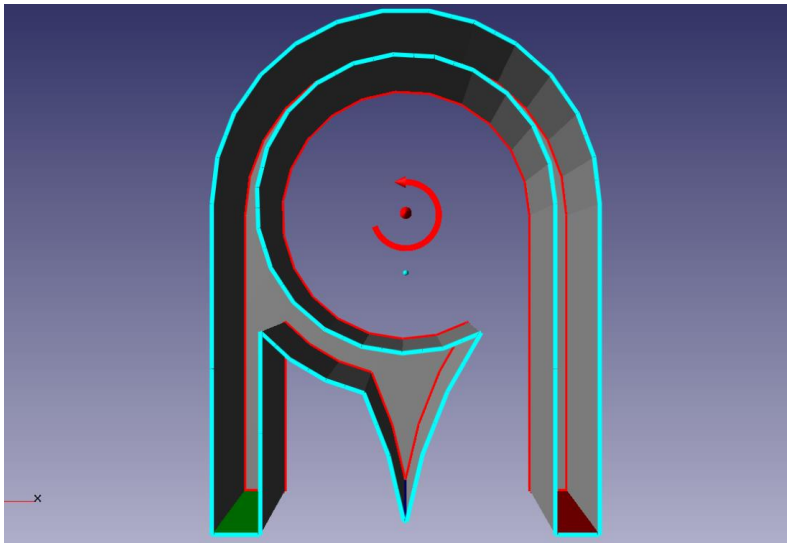


Figure 2: 3D model of an air purification device in a section with boundary restrictions

The device for purification of gas media contains a positive electrode and a negative electrode, the role of which is played by the device body. The positive electrode is made in the form of a flat cylindrical body and is mounted with the possibility of moving in the opposite direction of the gas supply relative to the fixed negative electrode (red circular arrow). The device is provided with a nozzle (green plane) for supplying the gas space to be cleaned into the space between the positive electrode and the negative electrode, as well as a nozzle (red plane) for discharging the cleaned gas space. The device contains a power supply, the output of which is connected to the input of a voltage rectifier, the output of which is connected to the input of a generator, the output of which is connected to the input of a voltage multiplier, the output of which is connected to the positive electrode. The device is equipped with a hopper for collecting particles removed from the gas space in the form of a cone (top is blue). The positive electrode is mounted on a platform mounted on the rotor of the drive mechanism (not shown). On the platform, on the inner surface of the electrode, there is a power supply, a voltage rectifier, a generator and a voltage multiplier. The device is equipped with a collection hopper for collecting particles deposited on the positive electrode.

The contaminated gas space is fed into the space between the negative electrode and the positive electrode, which rotates in the opposite direction to the gas space feed. The particles of the contaminated gas space, falling into the electric field created in the interelectrode space, are polarized and attracted to the outer surface of the positive electrode. The electrical circuit of the device, which consists of a power supply, a voltage rectifier, a generator and a voltage multiplier connected in series, creates a voltage of 1 60 kV or more on the positive electrode, which provides a high voltage of the electric field. The movement of the positive electrode relative to the fixed negative electrode towards the flow of the gas space to be cleaned, together with the high voltage of the electric field in the interelectrode space, significantly increases the efficiency of gas space cleaning.

The particles removed from the gas space under the action of their own weight fall into the hopper, which is subject to periodic cleaning.

3.2. Experimental validation of the model

To verify the theoretical positions, computer simulation was carried out using FlowVision software, the results of which are shown in Figs. 3–5. The simulation shows the movement of particles in the centrifuge in the form of vectors, complete fill and isolines (red – maximum values, blue – minimum or infinitesimal values).

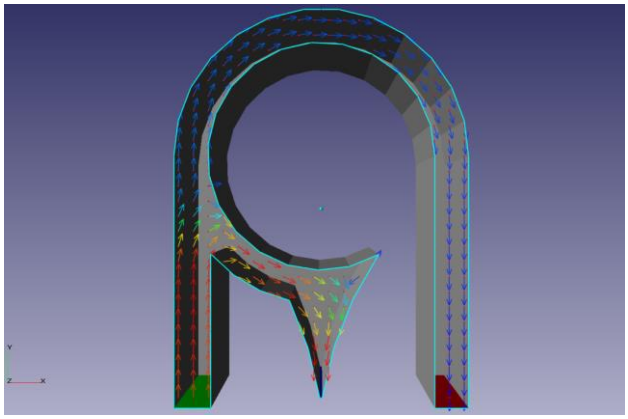


Figure 3. The velocity of particles in the air flow, displayed in vector form

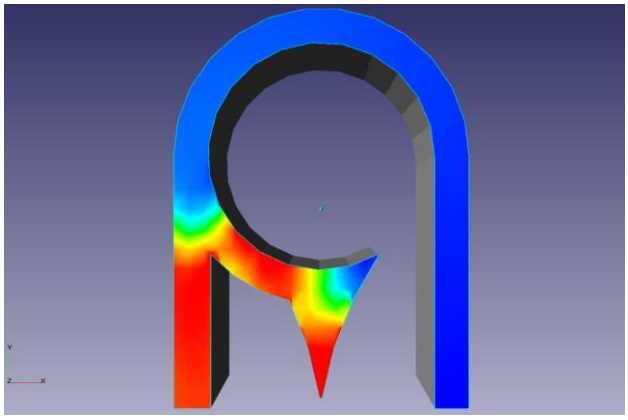


Figure 4. Concentration of particles in the air stream in the form of a complete flood

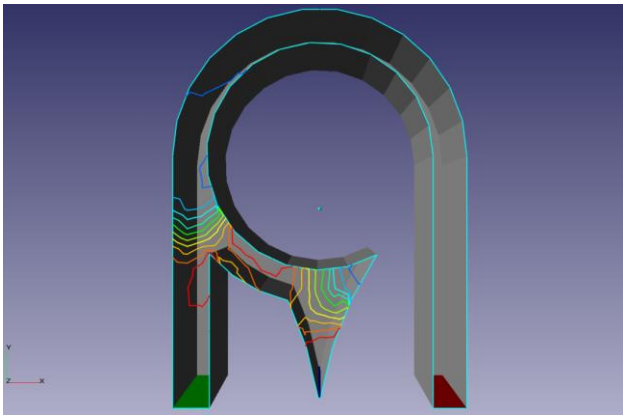


Figure 5. Concentration of particles in the air flow in the form of isolines

Based on the simulation results, it can be concluded that the flow of the gas space corresponds to the theoretically justified one and, therefore, can be used to clean the air from fine dust particles in the "modular type" plant described above, and, in combination with an additional electrostatic precipitator installed at the outlet, will ensure the final purification of compressed air streams and its subsequent ionization, since over time, individual dust particles that could not be separated by centrifugation remain in the air stream.

For example, there are many air treatment filters that are used in centralized and local ventilation systems for buildings and individual rooms. The vast majority of them are designed to clean the air from dust of a certain dispersion. The general disadvantage of such filters is their low efficiency, fixed level of air purification and changes in its ionic composition (deionization) due to electrification of filter materials (accumulation of static electricity).

In terms of air purification, an electrostatic filter is more efficient and flexible in use. The disadvantage of this type of filter is that it does not practically deionize the air completely. The most suitable is an air electrostatic filter with air ionization. This filter is the closest analogue and was chosen as a prototype. The main disadvantage of the filter is that, at an acceptable and controllable air purification efficiency, its deionization in the filter is compensated by ionization due to corona discharges. This leads to uncontrolled generation of harmful amounts of ozone (O_3) and nitrogen oxides ($NO)_x$.

The technical problem to be solved by this utility model is to preserve the ionisation of natural air and to regulate the efficiency of air purification (dispersion of the absorbed dust) without the use of electrostatic effects. The goal is achieved by using a polymer with a non-electrifying surface as a filter material and adjusting the air purification efficiency by changing the filter material's seal. Fluor plastic was chosen as the filter material because it does not electrify during operation (it is astatic).

The filter is constructed as follows: fluor plastic is placed in a cylindrical body consisting of two parts and a bolted connection. It is covered from above and below with a mesh of any mesh size for air to pass through. A top view and a longitudinal section of such a filter are shown in Fig. 6.

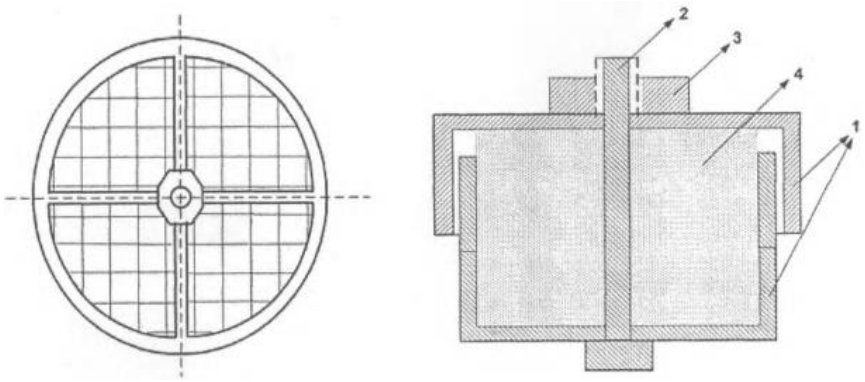


Figure 6. Electrostatic filter: top view and longitudinal section

The movable parts of the housing 1 hold the filter media 4 in the inner part by means of bolt 2 and nut 3. The change in the dispersion of the dust retained by this filter is regulated by compression of the filter material by means of the nut 3.

The use of computer simulation in the design of machines and mechanisms makes it possible to transfer the process of testing actually manufactured mechanisms to full-scale testing, which significantly saves material and time resources for the preparation and implementation of modern machines or mechanisms in production and guarantees their quality and reliability during operation.

4. Discussion and suggestions for future research

The generalized results of the study are important in several aspects that need to be highlighted. Firstly, the mathematical description of aerodynamic processes occurring in the dust collector made it possible to establish quantitative indicators of the key parameters of the functioning of the entire system, taking into account the geometry of the proposed device. Secondly, computer simulation made it possible to determine the velocity of particles in the dust collector and their concentration in its working area. It is worth noting that these data are fully consistent with theoretical data, which indicates that the proposed device is sufficiently efficient in air purification. Thirdly, the analysis of the patterns of particle concentration distribution showed that a certain amount of particles does not settle in the conical dust collector as a result of centrifugation, but continues to move in the air flow. This led to the recommendation to install an electrostatic filter at the outlet of the dust collector for the final purification of the respirable fraction of dust. The presented study complements the known data on the movement of particles in a two-phase flow in terms of a better understanding of the aerodynamics of the process. Thus, modern software tools allow us to obtain results that demonstrate full agreement with full-scale experiments, making it possible to implement technological systems by designing and testing both the geometry of the proposed devices and their operating conditions. Prospects for further research should be related to the assessment of the impact of all types of physical interactions, in particular, in combination with air ionization/deionization factors,

which will require additional calculation modules that take into account mathematical dependencies, which have not yet been fully derived.

5. Conclusions.

As a result of the comprehensive study, the following can be noted:

1) the developed computer model allowed us to establish the parameters, characteristics, requirements and limitations of a real dust removal device. The velocity of particles in the air flow was determined and displayed in vector form, as well as the concentration of particles in the air flow, which is displayed in the form of complete fill and isolines;

2) the flow of the gas space corresponds to the theoretically justified flow and, therefore, can be used to clean the air from respirable dust particles in a "modular type" installation;

3) over time, some dust particles that could not be separated by centrifugation remain in the air stream, so an additional electrostatic precipitator must be installed for final cleaning;

4) the assembly of cleaning devices in a "modular complex" allows to create a universal multi-stage plant that ensures complete dust and ash collection and gas neutralization. The sectional, "modular" layout reduces metal consumption, device cost and maintenance;

5) the practical implementation of the "modular complex" allows to increase the efficiency of separation and purification of flue and corrosive gases within a wide range of dust and gaseous pollutants, significantly reduce the degree of air pollution, as well as the cost of construction and maintenance of treatment facilities;

6) the use of the proposed device in comparison with all known analogues provides the possibility of effective cleaning of gas environments from industrial and household dust, as well as combustion products;

7) tests of the efficiency of the electrostatic precipitator using special equipment have shown that the level of air ionization after passing through such a filter practically does not change (within the error of the device), preserving its natural qualities. The electrostatic precipitator, when adjusted to the standard level, absorbs dust of any dispersion encountered in production conditions;

8) trial operation of the developed filter demonstrates its functionality, ease of manufacture, the ability to pass air in any direction with the same effect and the economic feasibility of using it in both centralized and local ventilation systems, buildings and premises.

All of the above makes it possible to make a general conclusion about the sustainable operation of the designed technological system, which can be recommended for implementation in the production sector to create a safe working area for employees.

References

Bai, Y. (2017). Grey Mathematics Model for Atmospheric Pollution Based on Numerical Simulation. *Chemical Engineering Transactions*, **71**, 679-684. https://doi.org/10.3303/CET1871114

Bayraktar, S. Turgut, Y. (2016). Investigation of the cutting forces and surface roughness in milling carbon fibre reinforced polymer composite material. *Materiali in Tehnologije*, **50** (2016) 4, 591-600. https://doi.org/10.17222/mit.2015.199.

Birkhoff, G. Hydrodynamics. (2015). Princeton University Press. URL: http://www.worldcat.org/isbn/9780691625911.

Biliaieva, V. V., Kirichenko, P. S., Berlov, O. V., Gabrinets, V. O., Horiachkin, V. M. (2019). Computer simulation of air pollution in case of dust cloud movement in open pit mine. Science and Transport Progress, (4(82), 18-25. https://doi.org/10.15802/stp2019/178556.

Chencheva, O., Lashko, Ye., Rieznik, D., Cheberyachko, Yu., Petrenko, I. (2023). Research of the aerodynamic process of carbon dust removal from the working zone. *Municipal Economy of Cities*, **1**(175), 208-220. https://doi.org/10.33042/2522-1809-2023-1-175-208-220.

Zhou, G., Liu, Y., Kong, Y., Hu, Y., Song, R., Tian, Y., Jia, X., Sun, B. (2022). Numerical analysis of dust pollution evolution law caused by ascensional/descensional ventilation in fully mechanised coal mining face based on DPM-DEM model. *Journal of Environmental Chemical Engineering*, **10**(3). https://doi.org/10.1016/j.jece.2022.107732.

Kumar, A., Schafrik, S. (2020). Multiphase CFD simulation and laboratory testing of a Vortecone for mining and industrial dust scrubbing applications. *Process Safety and Environmental Protection*, 144, 330-336. https://doi.org/10.1016/j.psep.2020.07.046.

Liu, Q., Nie, W., Hua, Y., Jia, L., Li, C., Ma, H., Wei, C., Liu, C., Zhou, W., Peng, H. (2019). A study on the dust control effect of the dust extraction system in TBM construction tunnels based on CFD computer simulation technology. *Advanced Powder Technology*, **30**(10), 2059-2075. https://doi.org/10.1016/j.apt.2019.06.019.

Salenko, A., Chencheva, O., Glukhova, V., Shchetynin, V., Budar, M. R. F., Klimenko, S., Lashko, E. (2020). Effect of slime and dust emission on micro-cutting when processing carbon-carbon composites. *Eastern-European Journal of Enterprise Technologies*, **3**(1 (105), 38-51. https://doi.org/10.15587/1729-4061.2020.203279.

Tannehill J.C., Anderson D.A. and Pletcher R.H.: Computational Fluid Mechanics and Heat Transfer (2nd ed.), Francis & Taylor, Philadelphia, (1997), 1-774.

Xiu, Z., Nie, W., Yan, J., Chen, D., Cai, P., Liu, Q., Du, T., & Yang, B. (2020). Numerical simulation study on dust pollution characteristics and optimal dust control air flow rates during coal mine production. *Journal of Cleaner Production*, 248, 119197. https://doi.org/10.1016/j.jclepro.2019.119197.

Received February 27, 2024, revised July 9, 2024, accepted August 27, 2024

Baltic J. Modern Computing, Vol. 12 (2024), No.3, 286-303
https://doi.org/10.22364/bjmc.2024.12.3.05

An Approach of ICT Incident Management Based on ITIL 4 Methodology Recommendations

Dalė DZEMYDIENĖ[1], Sigita TURSKIENĖ[1],
Irma ŠILEIKIENĖ[1,2]

[1]Institute of Regional Development Šiauliai Academy Vilnius University, Vilniaus str. 88, LT-76285 Šiauliai, Lithuania
[2]Department of Information Technologies Faculty of Fundamental Sciences Vilnius Gediminas Technical University -Vilnius Tech, Saulėtekio ave. 11, Vilnius, Lithuania

dale.dzemydiene@mif.vu.lt, sigita.turskiene@sa.vu.lt,
irma.sileikiene@vilniustech.lt

ORCID 0000-0003-1646-2720, ORCID 0000-0002-2019-6712,
ORCID 0000-0002-1185-0970

Abstract. The main goal of this research is to develop an approach for management of incidents of IT infrastructure library following the recommendations of the ITIL 4 methodology. This approach is provided for solving of ICT incidents more efficiently in an educational institution, focusing on the value creation of ICT services and their maintenance. When ICT infrastructure disruptions occur, ways and appropriate measures must be found to deal with incident management issues. The methods of recognition of the ICT incidents are included in decision support subsystem by applying classification and prioritization methods. The issues of choosing the right software in order to more effectively automate the management are proposed by analysing the process of solving emerging ICT infrastructure incidents. Functional capabilities of Spiceworks Help Desk software are explored to help in registration and management of the incident resolution cases caused by ICT disruptions. The article examines ITIL incident management practices and methods of solving ICT incidents in an educational institution, which become one of the most important in ensuring the smooth and uninterrupted work of interoperable systems of education institution.

Keywords: information technology infrastructure library **(**ITIL); information communication technologies (ICT); ICT incident management.

1. Introduction

The violations of information technology infrastructure library (ITIL) components are increasing nowadays and different types of ICT incidents can disrupt various links of infrastructure of ICT and the activity processes of services and their uninterrupted work in the institution. As the number of incidents in cyberspace increases, the problem of ensuring the smooth operations of ICT infrastructure arises in business enterprises, public sector and in educational institutions, where a consistent working process is carried out. This requires the continuous operation of the functions of interoperable

systems, provided by ICT infrastructure services (Axelos Limited, 2019; 2020; ITIL4 Practice guide, 2023). In an educational institution, it is important to ensure uninterrupted work of ICT chains and to effectively manage ICT services, especially focusing on incident management practices and technologies.

ITIL incidents refer to various ICT infrastructure disruptions and ICT service interruptions, such as, for example, disruption of the main server of the institution, power outages, computer network violations, violations of individual software modules, etc. and service performance degradation (Axelos Limited, 2021). Most often, incident management is integrated into the entire ICT service management process (Gillingham, 2023; Darby, 2022; Dzemydienė et al., 2022; Kaplan, 2023). The purpose of ITIL incident management is to ensure the timely restoration of services to normal working conditions by restoring damaged ICT provision services, minimizing the impact of the breach resulting from various disrupted chains (Dzemydienė et al., 2023).

The article presents ICT incident management recommendations based on the ITIL v4 methodology. The goal is to create an incident management and resolution algorithm that would enable effective incident management according to the set priorities. A method of incident prioritization is proposed, which is integrated into decision-making in the course of eliminating service disruptions. The research includes the analysis of computerized incident management systems and the application of these tools in solving cases of ICT disruption in an educational institution and managing their resolution process.

Spiceworks Help Desk software was chosen for solving ICT incidents and performing the main management steps, which is free, has an easy-to-use interface, has automated incident registration and management, network monitoring, report generation and the ability to integrate with other ICT management systems, integrated remote access to user device function.

Innovative ICT, such as service-oriented architecture, cloud technologies, application of templated scenarios, creation of open data access, provides opportunities for more intensive development of data exchange and reuse of data and increases the efficiency of services. The legal acts of the Republic of Lithuania and the directives of the European Union (EU) oblige to move to more effective forms of digitization and integrative possibilities of information systems (IS) (Lithuania's Progress Strategy "Lithuania 2030", 2012).

The legal and technical base in the public sector is sufficiently prepared, harmonized and meets EU requirements and standards for the use of official systems (Lithuania's Progress Strategy "Lithuania 2030", 2012). However, educational institutions are still hesitant to move to the innovations and opportunities of innovative tools. There is a sense of digital differentiation in the use of administrative tools. There is a lack of an integrated, coordinated approaches, which can enable the adaptable application of information resources, their implementation in the infrastructure of educational institutions at all levels.

The specificity of the management of educational institutions means that employees of all levels share the coordination of management, working groups and activity planning. Participating at each level, the main customers - teachers, middle managers, managers - are interested persons who strive for the implementation of common goals,

and their work principles are transferred to the operational processes of information systems (IS) that enable the development of automated systems. The benefits of ICT management in education have implications for better collaboration between the school as an administrative unit, parents and external institutions and local authorities. The management of ICT services influences the productivity and efficiency of the work of an organization, company or institution, the dependence on paper documents decreases, an organized information and service transmission system is created, which considers the needs of the institution (DESI, 2022). Institutions that do not implement ICT service management innovations have risk for losing of the ability to effectively manage complex processes.

The aim of this research is forwarded for development of approach of implementation of ITIL 4 methods for analysis and maintaining of the infrastructure of ICT of educational institution. The objectives of this research are concerning the development of constructional structure of description of activities of educational institution for developing of ICT library following to requirements of the ITIL 4 methodology. The set of recommendations are provided for analysis of ICT infrastructure management services by showing more possibilities of modern ITIL 4 methodology application tools. The objectives are realized by showing the possibilities through the solving some problems of ICT service management according to detecting of ICT incidents and realizing them. The results are related to the development of ICT functional capabilities for the modernization of the work of secondary education institution, by demonstrating the advantages of ICT management services. The results of experimental research enable to provide recommendations for selection of ICT infrastructure management tools and demonstrate their functional advantages. The experimental results with additional forms of integrated information systems (IS) helps to form a new understanding of value acquired through ICT infrastructure development and services efficiency. The article describes how the development possibilities of ICT management services are accesses, and how it is possible to implement by the certain methodology and tools, which become significant in the works of educational institution. The implementation strategy of methodology of ITIL 4 is recommended for the application for ICT infrastructure management services by realizing optimization of administrative processes in the educational institution.

The content of this article is separated in chapters. In 2 Chapter is presented the review of practices of IT incident management and describes main principles. The 3 Chapter presents the main stages of our proposed method for categorization of incidents during diagnostic stage and algorithm which is implemented into the management of ICT incidents. The 4 Chapter is devoted for representation of experimental research results of proposed methodology for management of ICT incidents in the infrastructure of concrete educational institution. In Conclusions we are summarized our obtained results and present the recommendations of ICT management for educational institutions.

2. Review of practices of management of ICT incidents and ITIL 4 methodology recommendations

The continue and not interrupted ICT work implies effectiveness of economy and work of business enterprises (Dzemydaitė and Naruševičius, 2023; Dzemydienė et al., 2022).

But we deal with the problem of ICT interruptions. In the past, detection of ICT incidents was mainly based on information from end users and ICT professionals.

Various ICT incident resolution methodologies are offered in scientific and practical activities (Palilingan and Batmetan, 2018; Danby, 2022; Thirhappa, 2023). However, it is the ITIL methodology and its latest version - ITIL 4 implies into the value of ICT and covers the most important aspects of ICT incident management (Axelos Limited, 2020; ITIL 4 Practice Guide, 2023). ICT incident management methodologies and specialist practices attempt to ensure that periods of unplanned service unavailability or degradation are kept to a minimum (Shepherd, 2019, Howells, 2020). This is made possible by two main factors: early detection of incidents and rapid restoration of their normal operation. ITIL v4 refers to incident management as a service management practice that describes the key activities, inputs, outputs and roles of those involved in the resolution process (Thirthappa, 2023; Key ITIL Concepts, 2023). Based on these guidelines, institutions are advised to create an incident management process that meets their specific requirements and operational specifics (ITIL 4 Practice Guide, 2023). Many incident management practices are divided into certain phases, with the key steps of incident management and periodic incident review important, and where each step is followed by an abstract sequence of actions (ITIL 4 Management Practices, 2023).

Modern good management practices suggest detecting and logging incidents as soon as they occur, before they affect users. Implementation of this method for ICT incidents management (Axelos Limited, 2021) has many advantages, according to:

- shortened duration of service unavailability or deterioration;
- higher quality raw data supports the correct response and resolution to incident resolution, including automatic inclusion of a resolution method, based on the resolution of previous analogous cases;
- some incidents remain invisible to users, thus improving user and customer satisfaction;
- some incidents can be resolved before they affect the agreed service quality of customers, improving perceived service and officially reported service quality;
- incident-related costs can be reduced.

The incident detection process is enabled by monitoring of conditions of ICT work, event management practices and implementing of right software for managing of incidents. Our proposed approach is based on the inclusion of event categorization method for diagnosis and management process that differentiate incidents from informational events and alerts.

Effective incident resolution can become a permanent way of dealing with significant issues in their aftermath. If the incident is not resolved in a timely manner, the issue remains in an error state and should be addressed through documentation when related incidents occur. Each documented solution should include a clear definition of the symptoms to which it applies. In some cases, the solution to the ICT incident can be automated. For other incidents, you need to find a way to fix the error. This is part of error control. Error control activities manage known errors, which are issues for which initial analysis has been completed; this usually means that faulty components have been identified. Error control also includes the identification of potential permanent decisions

that may be subject to a change request, but only if this can be justified by the costs, risks and benefits.

Table 1. List of components of ITIL implemented in X Educational institution

The ICT infrastructure of X Educational institution						
No	**Activities/ functions**	**Hardware and Computer networks**			**Systematic standardized software**	**Applicational software**
		Types of Networks (WAN, LAN, WIFI, INT)	**Computer Work Stations**	**Other Tech-nology**	**Operation systems. Cyber Security systems**	**Application Software Systems and Web Tools**
1.	Management of information and communication	WAN, LAN, INT (until 1 Gb/s, WIFI (~ 1 Mb/s, -~200 Mb/s). INT - fiber optic connection through LITNET	1. Ethernet 144 and more stationary working educational and teaching stations. 2. Wi-Fi – ~600 laptops and smart devices.	Software of Internet control: 1. Ethernet routers and switches, ~ 4 controlled and ~ 6 not controlled; 2. Wi-Fi access through MIKROTIK RB760iGS routers and switches UNIFI US24P250, 26 Wi-Fi points UNIFI U7LR.	Operation systems: 1. Stationary work station OS – MS Windows 10/11. 2. Smart devices with OS Android.; 3. Smart devices with iOS. Security systems: 1. In stationary work and educational stations – standard MS Windows OS safety toolkits 2. LITNET supported Ethernet and Wi-Fi Internet access data monitoring and filtering system using Fortigate with UTM software.	LITNET provides a monitoring and filtering system for Ethernet and Wi-Fi connection to the Internet Arrangement of Internet explorer systems and interoperable connection software for Data bases and data warehouses
1. 1	Internet and cloud for communication and information sharing	WAN, LAN, INT (~ 1 Gb/s, WIFI (~ 1 Mb/s, ~200 Mb/s). INT fiber optic connection via LITNET	All PCs and smart devices		In „Google For Education platform applied safety and filtering systems	The Google For Education platform provides applications for stationary work and learning places and smart apps for phones, tablets and smart screens. Office 365 platform provided to schools by emokykla.lt. The official page of the organization - program PyroCMS, Inc 0.17 s | 14 mb v3.3.3
1. 2	Management of web site of the institution		All jobs and Cloud service stations		Only employees authorized by the institution can connect to the content management systems (CMS) and domain management system of the website.	1. Web platform for managing of web site of educational institution;2. the internet system: iv.lt -web client system for internet site management and domain control and internet service support plan realization.3. The system wolet.lt - virtual server services for web site hosting

The effectiveness of incident solutions should be evaluated each time a solution is used, as the solution can be improved based on the evaluation (Palilingana and Batmetan, 2018). Problem management activities are very closely related to incident management. Practices must be designed to work together in the value chain of organization. The activities of these two practices can be complementary (for example, determining the cause of an incident is a problem management activity that can lead to incident resolution), but they can also conflict (for example, investigating the cause of an incident can delay the actions needed to restore service). Examples of links between problem management, risk management, change enablement, knowledge management and continuous improvement include.

ICT specialists, which are responsible for diagnosing problems, often need the ability to understand complex systems and think about - how various failures could have occurred. Cultivating this combination of analytical and creative skills requires mentorship and time, as well as appropriate training. For these needs applicable became the construction of ITIL as componential and multi-layered infrastructure library, from which the ICT specialists can decide about destroyed components and their relationship with other software components and their dependencies. The part of such components of ITIL in concrete X Educational institution is presented by their relationships in Table 1. All components of ICT library are analysed in our previous work (Dzemydienė et al., 2023).

There are important to extract the set of activities, which are applied in the ICT incident management practices. The activities, which are related to ICT incident management and highlighted as the most important and described in the practice recommendations are shown in Table 2.

Table 2. Different incident management activities related to ICT incident management practices

Activity	Related practices
Investigating the causes of ICT incidents	Such works are solving in problem management stage
Communication with users	Activities in responsibility of ICT Service department
Implementation of changes to products and services	Such works are solved in stages of change enablement; deployment management; project management; release management; software development management
Monitoring the activities of technologies, teams and suppliers	Continues monitoring and event management with software
Management of improvement initiatives	Continuous improvement
Management and fulfilment of service requests	Management of service requests implement the Help desk software
Restoring normal operations after a disaster	Responsible the Service Continuity Management

As the ITIL 4 methodology is universal and flexible, its recommendations only indicate the direction and not specific solutions.

3. The approach of formalization of the decision support activities for solving ICT incident management problems

Problem management activities can be organized as a specific case of risk management: they aim to identify, assess and control risks in any of the four dimensions of service management. It is useful to adopt risk management tools and techniques for problem management. Implementing a solution to a problem is often outside the scope of problem management.

Issue management typically initiates resolution through change enablement and participates in post-implementation review and approving and implementing changes (Figure 1).

Therefore, the ITIL4 incident management flow chart (Figure 1) has 6 steps, which often do not answer all the questions that arise. Therefore, according to the specified direction, it is recommended to create sequences of decisions, actions and processes that meet the needs of your institution.

The algorithm allows actions to be taken when a certain incident occurs and specifies a sequence of actions to resolve the incident. Since ITIL 4 recommends describing the processes by adapting them to your organization, the incident management process of a specific organization can change significantly. In addition, problem management can use knowledge management system information to investigate, diagnose, and resolve problems. Problem management activities can identify opportunities for improvement in all four dimensions of service management. In some cases, solutions can be treated as improvement opportunities, so they are entered into a continuous improvement register (CIR) and continuous improvement methods are used to prioritize and manage them, sometimes as part of backlog products. Many problem management activities rely on employee knowledge and experience rather than following detailed procedures.

The component-based ICT infrastructure of education institutions is shown in Figure 2. It shows that a modern school has a substantial ICT infrastructure, including managed and used information systems, IT services and hardware (here, LMS – Learning Management System, LDAP/AD -LDAP or Active Directory; DMS- Document Management System; CMS- Content Management System, SMTP – e-mail server). This IT infrastructure is becoming difficult for schools to manage, and therefore requires a centralized Service Desk and software to manage it efficiently and respond quickly to incidents.

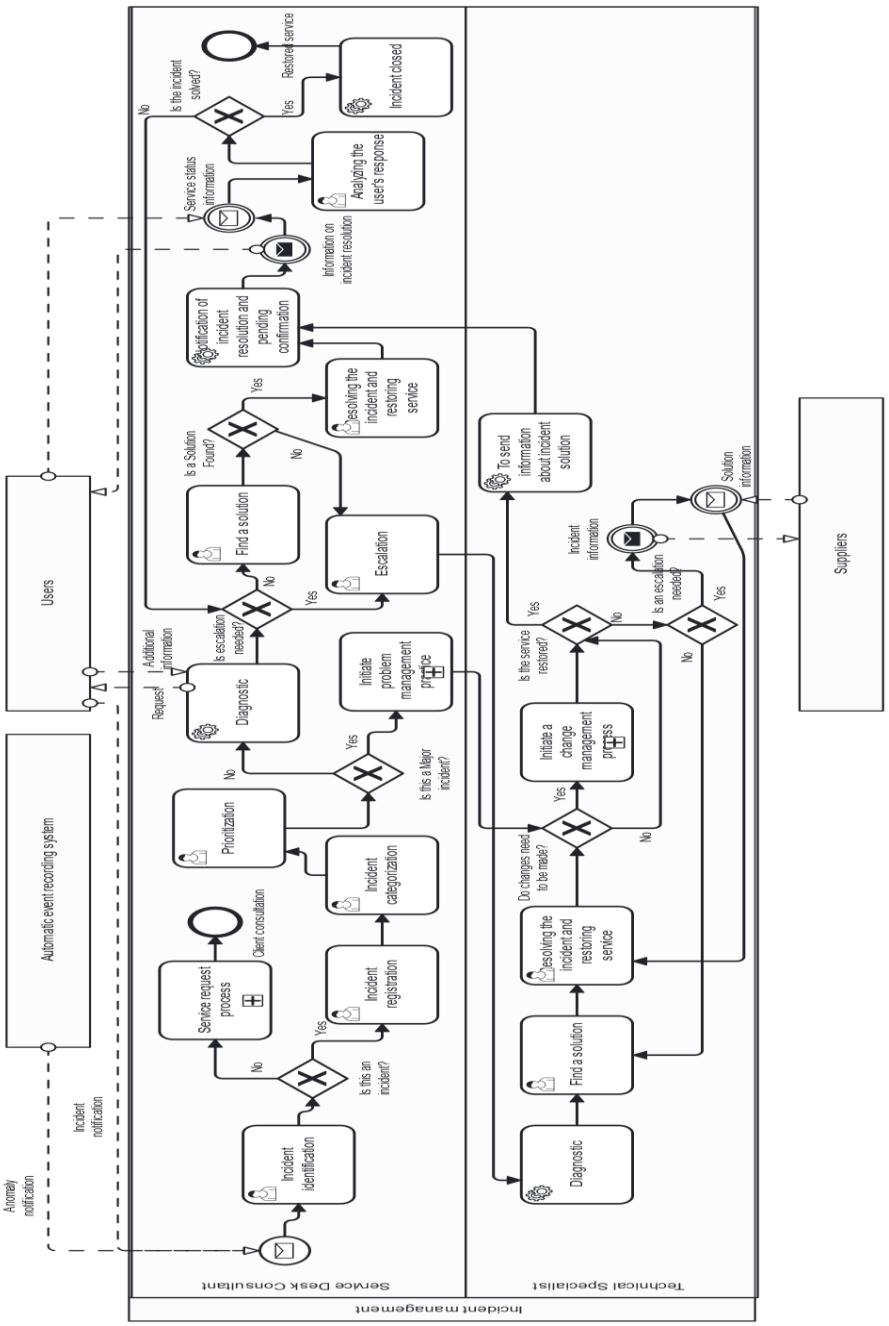


Figure 1. The algorithm of processes of management (including recognition and categorization stages) of ICT incidents (designed by using BPML notation)

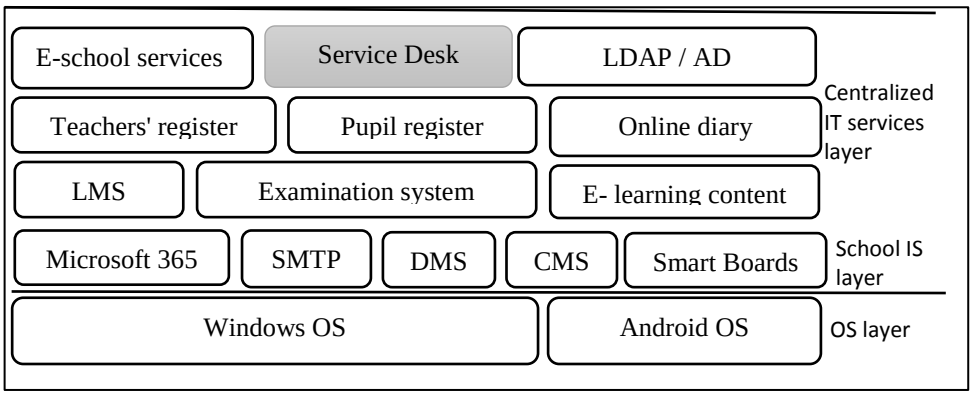


Figure 2. Component-based IT infrastructure of the educational institution

To deal with one of the problems of evaluation of priority of incidents of ICT infrastructure we propose such decision support model in which is introduced the utility function for the different kind of messages about concrete incident. Such messages are evaluated and stored in the weighted context matrix (M_L) for every activated l message (m) about incidents from n layer of ICT infrastructure.

$$M_L = \begin{pmatrix} d_{L_{11}} & d_{L_{12}} & \dots & d_{L_{1n}} \\ d_{L_{21}} & d_{L_{22}} & \dots & d_{L_{2n}} \\ \dots & \dots & \dots & \dots \\ d_{L_{l1}} & d_{L_{l2}} & \dots & d_{L_{ln}} \end{pmatrix} \quad (1)$$

The utility of the messages about incidents can be weighted in a function which assigns a value to each message about incident to be disseminated. The value is calculated by the equation (2):

$$d_{L_{ij}} = \left(Ty_j + H_j + Ex_j\right) m_i cr_i Pr_i, i = 1,...,l,\ j = 1,...,n \quad (2)$$

where Ty is the type of destroyed data which values are from the section [1;3]. The values form such section are obtained as follows: (1— the incident is evaluated as not so much important, 2—important, 3—very important). H is the parameter in the interval [0, 1] showing two possibilities:

if the data should be used for historical saving (1) or not (0). Ex is the parameter in the interval [1–4] showing the destroying software infrastructural level (1— for level of Operation systems; 2— for level of IS and own administration systems; 3— for level of other IT services; 4— for level of centrally administrated systems and services) and cr is the coordinates of the software location.

The priority of the message (Pr) is calculated by formula: $\mathrm{Pr}_j = 1 + \frac{\mathrm{I_j}}{\mathrm{A_j}}$ and it is normalized with values falling in a predetermined interval [1–3], where 3 means that the message about the incident priority is critical and it must be disseminated immediately, 2 means that the message about the incident have medium priority, and 1 means that the message about the incident is not important and can be suspended or rejected. $\mathrm{I_j}$ is importance of the message about incident in the interval [0, 1] where 0 is when very high importance is related message and 1 is not so much importance related message. $\mathrm{A_j}$ is message age function with normalized values in predetermined interval [1, 2, 3] which is calculated by (3) where $\mathrm{T_M}$ is difference between current and message compilation time.

$$A = \begin{cases} 1, \text{if } \mathrm{T_M} > 5s \\ 2, \text{if } 1 < \mathrm{T_M} < 5s \\ 3 \text{ if } \mathrm{T_M} < 1s \end{cases} \tag{3}$$

The predictive utility of the incident message is based on assigning of weights by a function that assigns a value to each message, enabling to recognize the priority of incidents and to be passed to the recipient entity. The value is calculated according to the formula (2).

According to ITIL recommendations, after registering an incident, it is suggested to move to its categorization stage (Figure 2). This phase consists of a 2-part incident categorization and prioritization. During the categorization phase, incidents are grouped by importance and complexity. Greater granularity is also possible to facilitate their management and analysis (Table 3).

Categorization helps isolate and group incidents based on their nature, such as software errors, equipment failures, or service disruptions. It is important to prioritize disruptions, considering the impact of the incident on the operation of the educational institution and its urgency.

In the prioritization phase, incidents are analysed according to impact and urgency and their levels in order to manage and resolve them more effectively (Table 3).

Priorities are divided according to the impact of the incident on the operation of the institution and according to how quickly the incident needs to be resolved. The impact level indicates how strongly the service disruption affects the organization's operations and the organization's users. Urgency indicates a measure how long it will be until an incident has a significant impact on the business. It is important to decide how long an incident affected the service itself, whether it is completely disrupted and cannot be used at all, whether the service can be used partially, etc. Each organization should develop descriptions of impact, urgency and priorities based on its activities.

Table 3. Categories for assignment of priorities for importance of incidents

		Impact on the disruption of the institution activities			
	Priority	Critical	Tall	Average	Low
Priority according to urgency	Critical	1	1	2	3
	Tall	1	2	3	3
	Average	2	3	3	4
	Low	3	3	4	5

5-level priorities are most commonly used according to (Danby, 2022):

Priority 1 - awarded when an ITIL incident resolution response can be provided within 10 minutes and incident resolution can take up to 3 hours.

Priority 2 - is given when an ITIL incident resolution response can be provided within 20 minutes and the incident resolution duration is up to 6 hours.

Priority 3 - given when an ITIL incident resolution response can be provided within 1 hour, incident resolution duration up to 2 working days.

Priority 4 - given when an ITIL incident resolution response can be provided within 5 hours, incident resolution duration 5 working days.

Priority 5 - given when an ITIL incident resolution response can be provided within 1 day, incident resolution duration up to 2 weeks.

When dealing with incidents, it is very important to assess whether the incident may affect other areas of a measure how long it will be until an incident has a significant impact on the business ICT. During the elimination of the consequences of the incident, the possibility of the spread of the incident and the possible impact on other areas of ICT are investigated.

After eliminating the incident, it is necessary to document it, i.e., to describe what incident happened and to supplement the knowledge base with the methods of solving it. The knowledge base must describe the resolution of the incident related to the signs of the occurrence of the incident.

4. Experimental research of the approach for ICT incident management in an educational institution

According to the recommendations of the ITIL 4 methodology, it is important to properly create a knowledge base covering as many typical procedures as possible, which describes the execution processes and the occurrence of certain incidents and their resolution methods. An educational institution needs to implement standardized cyber security measures and it is important to follow them. The incident management process is considered successful, the result of which is the resolution of the incident in a fast and optimal way, minimizing the impact on service users. The institution, considering its ICT infrastructure and services provided, must create clear and effective incident resolution procedures.

During the experimental study, incidents in the educational institution were recorded using the Help Desk system. Before recording incidents, it is necessary to describe exactly what will be considered an incident. Axelos (2019) compiled guidelines that enable an incident recognition scheme (Figure 3) that indicates when an event can be recorded as an incident and when an event is not considered an incident.

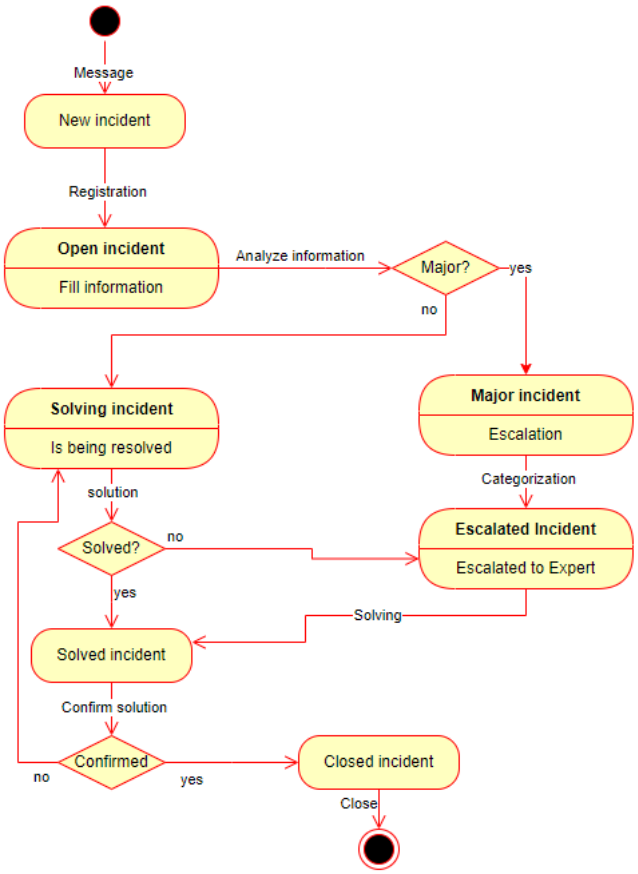


Figure 3. The algorithm of ICT incident solving

Once an event is determined to be classified as an incident, it is logged in the Help Desk system. Incident management requires a clear description of the process. In the ITIL incident management diagram (Figure 3), it can be seen that incident management is divided into three parts. The first part is registration in the Help Desk system, after which registered incidents are categorized and prioritized. After the incident is resolved, a report and response are prepared to responsible staff and users about the complete resolution of the incident.

If it is determined that it is an incident that requires specialized assistance for its solution and the defined category and assigned priority, the management of the incident is transferred to the institution's ICT technical support. After investigating the incident and reviewing the available knowledge base article, a solution is found to restore the service and, if necessary, additional knowledge base article is added. If the incident cannot be resolved, it is transferred to other specialists, for example for internet providers, technical support of various registers, administrators of e-school accounts, administrators of e-services, etc.

Because ITIL 4 is focused on the user and value creation, almost all events are treated as incidents. In the educational institution, it was decided to follow such an approach that any disruption of ICT services during which the user faces disruptions and inconveniences of disruption of ICT services will be considered as an incident. The scheme of the incident recognition algorithm (Figure 3) was used as the main tool for deciding the type of incidents that occur.

Categories whose violations affect priority assignment can be:

- Software failure;
- Network failure;
- Printer failure;
- Computer system software failure;
- Failures of projectors and smart screens;
- Email mail failures;
- Other equipment failures.

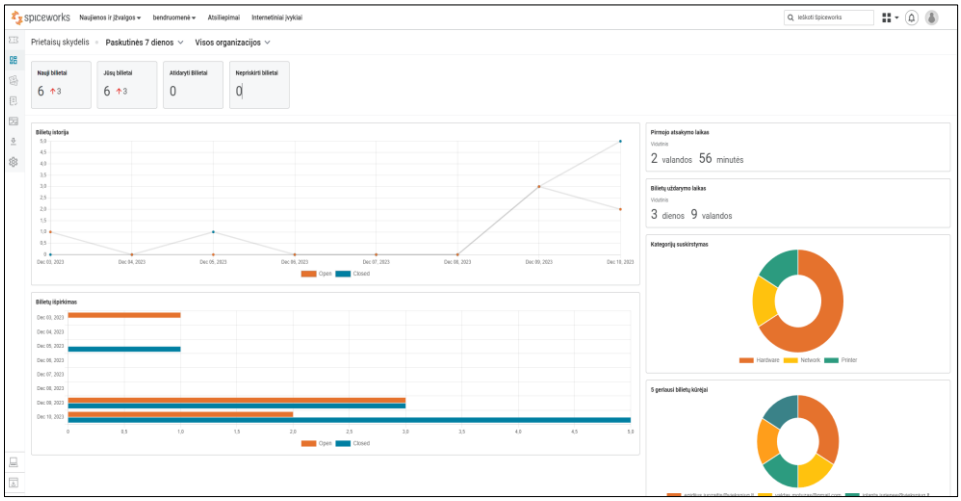


Figure 4. The monitoring of ICT incidents in Help Desk – Spice work package

However, within the institution, the teams involved in incident resolution are limited in resources and are often involved in other types of work at the same time. Some incidents should be prioritized over others to minimize the negative impact on users and optimize the use of resources (Axelos Limited, 2019).

In the educational institution, it is recommended to give priorities according to 3 levels:

• High-level - response to ICT incident resolution is possible within 15 minutes, incident resolution time up to 2 hours.

• Medium level - response is possible within 30 minutes, incident resolution time up to 8 hours.

• Low level - response is possible within 1 hour, incident resolution time up to 3 days.

After creating the priority matrix, it is necessary to determine the feedback and solution time intervals for each level. The organization must clearly define which incidents are high-level and which are low-level priorities based on urgency and impact. There can be no ambiguity. It is recommended to divide the level of impact of the incident according to how many users experience the inconvenience caused by the incident.

Possible categorization of incidents and examples of incidents by priority level:

• High level - failure of the institution's server, failures of the main network switches or routers. Those failures affect a large group of workers.

• Medium level - failure of the computer in the training class, Internet problems in one class. These are equipment failures or ICT service disruptions that affect a very small number of users, but have a significant impact on their operations.

• Low level - network printer failure, institution library computer failure. These failures do not significantly affect the educational process of the institution.

The Spiceworks system provides an opportunity to solve different types of faults (Figure 5). This system can help not only when administering several departments of the institution, but also when you want to test new configuration settings or changes and you don't want to distort the data in actual work.

During the entire 3-month incident registration period, 22 potential incidents were reported, of which 18 were reported by phone, 3 were reported by email, and one was reported through the incident reporting portal. During the experiment, it turned out that registration by phone is the easiest and most reliable method for employees of the institution.

Of the 22 reports of potential incidents received, 9 were false incidents. A false incident can be defined as the inability to properly use ICT services due to improper user actions.

Of the 13 registered incidents, their failure categories were distributed as follows:

• 6 network failures;

• 3 printer failures;

• 2 computer equipment failures;

• 1 failure of projectors and smart screens.

• 1 email failure.

The registered incidents were divided according to priorities: High priority - 5, Medium priority - 5 and Low priority - 3. The percentage of high priority incidents is unusually high, because the network failures affected a large part of the users, so the solutions to the incident had to be implemented quickly.

Incident management metrics are an important part of ICT service management as they help organizations monitor how effectively ICT disruptions and incidents are handled. Metrics can cover a variety of aspects, from the number of incidents to the time it takes to resolve them. Table 4 provides key practice metrics that should be applied depending on the institution's context, such as incident priority levels, expected incident resolution periods.

From the list of key incident management evaluation parameters, it follows that the weakest point is in automation. Both the early detection of incidents and their automatic resolution are not properly included in the institution's incident management. There is no clear indication of user satisfaction with incident management and resolution, as there is no integrated assessment tool in incident management reports. Incident management and resolution times do not exceed the times assigned to the priority levels.

A maturity model is used to check how one of the good practices can be applied in the ICT management of the institution. The ITIL Maturity Model defines the following competency levels for any management practice:

Table 4. Main parameters of incident removal detected during the experimental 3 months period

Practiced success factors	Main evaluation parameters	Results
Detect incidents early	Time from incident occurrence to detection.	Δt is 1 working day period
	Percentage of incidents detected by event monitoring and management.	0%
Fast and efficient resolution of incidents	The time from the detection of the incident to the start of the diagnosis.	High level, when $\Delta t < 11$ min.
	Time of diagnosis.	Intermediate level, when $\Delta t <$ 25 min.
	Number of assignment changes.	Low level, when $\Delta t < 1$ hour
	The percentage of waiting time in the total incident management time	High level , when $\Delta t < 15$ min.
	First time solution frequency.	Intermediate level, when $\Delta t <$ 30 min.
	Fulfillment of the agreed solution time.	Low level, when $\Delta t < 45$ hrs.
	User satisfaction with incident management and resolution.	Is expressed in grade [0;5]
	Percentage of incident that was resolved automatically.	0%
	Percentage of incidents resolved before notifying users.	77%
Continuous improvement of incident management	Incident resolution percentage using previously identified and recorded solutions.	100%
	Percentage of incidents resolved using incident patterns.	There are no exact data
	Improving key practice indicators over time.	0%
	Balance between incident resolution speed and efficiency metrics.	0%

Δt – is time duration spending on the process

Level 1. The practice is not well organized; it is executed as initial or intuitive. It may occasionally or partially achieve its goal through an incomplete set of activities.

Level 2. The practice systematically achieves its goal through a core set of activities supported by specialized resources.

Level 3. The practice is well defined and achieves its purpose in an organized manner, using dedicated resources and relying on inputs from other practices that are integrated into the ICT service management system.

Level 4. The practice achieves its goal in a highly organized manner, and its results are continuously measured and evaluated in the context of the service management system.

Level 5. Practice continuously improves organizational skills related to its purpose.

For each practice, the ITIL Maturity Model defines criteria for each capability level from level two to level five. According to these criteria, the practice's ability to fulfil its purpose and contribute to the institution's ICT service value system can be assessed (Axelos Limited, 2019).

The most important aspects of implementing ITIL incident management practices in an educational institution according to (Thirthappa, 2023) are:

• Inclusion of regular instructions for user's instructions: clear instructions must be provided on how to report incidents and what to do in the event of an incident.

• Organizing the feedbacks: is necessary to regularly collect feedback from users and adjust processes based on the information received.

• Applying the flexibility of adaptation of alternative software during incident's solving process: IT specialists have to be able to adapt to changing needs and circumstances in the educational environment.

• Forming of innovative organizational ICT culture: in ICT culture that values openness and learning, employees are more likely to actively participate in incident management processes, share ideas and learn from incidents. The culture encourages a proactive approach to incident management, emphasizing learning and improvement rather than assigning blame.

• Organizing of responsible employee training: provides employees with the necessary skills and knowledge to effectively respond to incidents, helps to better understand the incident management process. In a culture that encourages learning and development, employees are better equipped to handle incidents, allowing for faster and more effective decisions.

The involvement and commitment of managers and leaders promotes a fair and effective incident management culture. They can invite open communication, support learning opportunities, and lead by example.

Conclusions

An approach of ICT incident management is proposed in this article. The approach includes all stages of ICT incident management: from the ICT service management analysis of the educational institution, until identifying of the weak links for proper incident management. Among the weaker ones, we can mention the unforeseen procedures that must be solved by ICT employees, in the event of technical disturbances,

the failure registration and decision-making process is not carried out, the training of employees is not carried out systematically or not at all, attention is not paid to the creation of the institution's ICT culture.

In order to solve information technology infrastructure incidents, we apply the ICT incident management practices recommended in the ITIL 4 methodology, providing for specific steps to eliminate incidents, assigning responsible persons and choosing appropriate software tools for incident management, and training employees on how to react to incidents of the appropriate type.

Based on best practice metrics, it is recommended to create a knowledge base to store knowledge about incident resolution actions.

The experimental study showed that it is often difficult to determine the time interval between the occurrence of an incident and its registration, because incidents are registered only when users encounter disruptions in ICT services. Monitoring and event management practices, monitoring systems, and incident management software were used to address this issue, which enabled easier identification of incident types and immediate initiation of the resolution process.

Based on the data of the conducted research in accordance with the recommendations of the ITIL 4 methodology, it is proposed to organize incident management in an educational institution starting with incident registration, classification, decision-making, and solving the incident. Then the preparation of reports and the expansion of the knowledge base are prepared. Regularly review and improve incident management processes based on research and experience. This includes updating the processes, tools and methods used. Promote cooperation with other educational institutions and ITIL experts, sharing best practices.

References

AXELOS Limited (2019). ITIL Foundation: ITIL v4 Edition. ISBN: 9780113316069.

AXELOS Limited (2020). ITIL 4: Create, Deliver and Support: First edition 2020. ISBN: 9780113316328.

AXELOS Limited (2021). ITIL® Practices in 2000 words: Incident management, service desk, and service request management.

Danby, S. (2022). ITIL Priority Matrix: How to Use it for Incident, Problem, Service Request, and Change Management. Access by Internet [2023-09-10]: https://blog.invgate.com/itil-priority-matrix

Dzemydaitė, G., Naruševičius, L. (2023). Exploring efficiency growth of advanced technology-generating sectors in the European Union: a stochastic frontier analysis. Journal of Business Economics and Management, 24(6), 976-995. https://doi.org/10.3846/jbem.2023.20688

Dzemydienė, D., Dzemydaitė, G., Gopisetti, D. (2022). Application of multicriteria decision aid for evaluation of ICT usage in business. Central European Journal of Operations Research, 30(1), 323-343. https://doi.org/10.1007/s10100-020-00691-9

Dzemydienė, D., Turskienė, S., Šileikienė, I., Baltrukaitis, A., Kazlauskienė, A. (2022). Development of ICT infrastructure management services for educational establishments (in Lithuanian). ALTA'22. Advance learning Technologies and Applications. Annual International conference for education : Conference proceedings 30th of November, 2022 / edited by Danguolė Rutkauskienė. Kaunas: Kaunas University of Technology. p. 125-147. https://ndma.lt/alta2022/wp-content/uploads/2023/04/ALTA'22%20proceedings.pdf

Dzemydienė, D., Turskienė, S., Šileikienė, I. (2023). Development of ICT infrastructure management services for optimization of administration of educational institution activities

by using ITIL-v4. Baltic Journal of Modern Computing, vol. 11, no. 4, p. 558-579. DOI: 10.22364/bjmc.2023.11.4.03.

Gillingham, J. (2023). Key ITIL Concepts That One Should Know. Access by Internet [2023-10-22]: https://www.invensislearning.com/blog/key-itil-concepts/

Howells, C. (2020). ITIL Practices. Access by Internet . [2023-08-14]: https://cdn.ymaws.com/www.itsmfusa.org/resource/resmgr/ITIL4_Session_4-ITIL_Practic.pdf

ITIL®4 Practice Guide (2023), Incident Management. Access by Internet [2023-08-14]: https://www.scribd.com/document/600322911/Practice-Incident-management-ITILv4

ITIL®4 Management Practices. Access by Internet [2023-08-11]: https://www.knowledgehut.com/tutorials/itil4-tutorial/itil-management-practices-processes

Kaplan, S. (2023). What Is Organizational Culture and Why Is It Important? Access by Internet [2023-08-14]: https://www.psychologytoday.com/us/blog/the-power-of-experience/202312/what-is-organizational-culture-and-why-is-it-important

Key ITIL Concepts That One Should Know. Access by Internet [2023-09-17]: https://www.invensislearning.com/blog/key-itil-concepts/

Palilingan, V.R.; Batmetan J.R.(2018). Incident Management in Academic Information System using ITIL Framework. IOP Conference Series: Materials Science and Engineering, Volume 306, 2nd International Conference on Innovation in Engineering and Vocational Education,25–26 October 2017, Manado, Indonesia IOP Conf. Ser.: Mater. Sci. Eng. 306, 012110, DOI 10.1088/1757-899X/306/1/012110

Shepherd, H. (2019). ITIL 3 vs. ITIL 4 – What has changed and what is new? Access by Internet [2023-09-16]: https://advisera.com/20000academy/blog/2019/07/04/itil-3-vs-itil-4-what-has-changed-and-what-is-new/.

Thirthappa, K. (2023). Building a culture of incident response. Access by Internet [2023-11-12]: https://spike.sh/blog/building-a-culture-of-incident-response.

Received June 18, 2024, revised August 30, 2014, accepted September 13, 2024

Baltic J. Modern Computing, Vol. 12 (2024), No.3, 304-326
https://doi.org/10.22364/bjmc.2024.12.3.06

EnE-Rep: An Energy-Efficient Data Replication Strategy for Clouds

Mohammed ALGHOBIRI

Business Informatics Department
King Khalid University, Abha 62585 Saudi Arabia

maalghobiri@kku.edu.sa

ORCID 0000-0002-6414-739X

Abstract. The rapid rise in the popularity of cloud computing can be attributed to its inherent advantages. However, its expanding infrastructure leads to higher energy consumption and increased network latency. Virtual machine consolidation (VMC) and dynamic power management (DPM) are popular methods to improve energy efficiency. However, these energy-saving approaches are incompatible with data replication. Our approach in this study is called EnE-Rep, that categorizes cloud data center nodes based on workload and applies targeted strategies for each category. In addition, EnE-Rep leverages a robust collection of components including a load manager, energy monitor, and replicator for achieving energy-efficient data replication. Furthermore, intelligent placement decisions are made based on key factors like CPU utilization, server proximity, available bandwidth, and memory usage. Finally, CloudSim simulations validate the effectiveness of EnE-Rep, demonstrating significant reductions in energy consumption alongside improved performance metrics such as VM migration frequency, host shutdown rate, and data access time.

Keywords: Carbon Emission, Cloud Computing, Energy Efficiency, Data replication, Data center, Data-Intensive Computing.

1. Introduction

The rapid proliferation of cloud computing within the contemporary technological landscape can be attributed to its inherent advantages such as its ability to leverage a pool of shared resources that are readily scalable to meet user demands (Balakrishnan et al., 2017), Ruan et al., 2013). Cloud networks function by aggregating heterogeneous computing nodes from diverse locations. These nodes are then dynamically provisioned to users on an as-needed basis, offering a flexible and cost-effective solution. Service Level Agreements (SLAs) further govern the specifics of these cloud services, outlining performance guarantees and resource allocation. This 'on-demand scalability', a key feature of cloud computing, allows users to dynamically adjust resource utilization based on their evolving needs (Ding et al., 2015). Consequently, the scalability feature is

further bolstered by the "pay-as-you-go" pricing model, a revolutionary aspect of cloud computing (Buyya, 2009). By eliminating the need for upfront hardware and software investments, cloud computing empowers businesses to streamline operations and dedicate resources to core competencies (Beloglazov et al., 2012). Furthermore, economies of scale achieved through shared infrastructure contribute to the significant cost-effectiveness that drives widespread adoption of cloud computing solutions.

According to a report by CISCO, a staggering 94% of the total workload was processed by cloud computing in 2021. This widespread adoption can be attributed primarily to the ability of cloud infrastructure to provide access from anywhere in the world. Consequently, by 2020, a significant portion, 67%, of enterprise infrastructure had shifted to the cloud. Furthermore, cloud computing spending has grown at a remarkable pace, outpacing overall IT spending by a factor of six between 2015 and 2020. Currently, more than half of all IT spending goes towards cloud computing solutions (Kappelman et al., 2022). However, while the dynamism and flexibility of the cloud have undoubtedly fueled its growth, these features also present challenges, particularly regarding resource management, scheduling, and energy consumption (Jennings and Stadler, 2015), Ksentini et al., 2014). By 2020, cloud data center (DC) energy consumption was projected to reach an alarming 140B KW/H annually, equivalent to the energy produced by approximately 50 power plants. The financial and environmental costs associated with this immense energy consumption are significant. Quantifying this impact, the annual financial cost has reached $13 billion whereas the environmental cost translates to 100 million metric tons of CO2 emissions (Mytton, 2020). This high energy consumption is further reflected in global data center usage, which accounts for an estimated 205-Terawatt hours of power consumption per year which represents a significant 1% of the world's total energy consumption (Bonzi, 2021). The environmental impact is further emphasized by the fact that in 2018, data centers were responsible for a substantial 900 billion kilograms of carbon emissions, releasing approximately 4.4 kilograms of CO2 every hour (Bonzi, 2021).

Beyond scalability, another key challenge for cloud computing is efficient resource management, which directly impacts cost-effectiveness. In United States alone, estimates suggest that there are nearly three million data centers (DCs) accommodating approximately 12 million servers (Jahangir et al., 2021). However, it is worth noting that up to 30% of these servers are deemed unnecessary, with many others being underutilized. Despite this redundancy, the collective power consumption of these servers amounts to a substantial 140 billion KW/h annually, contributing to an alarming 150 million metric tons of carbon emissions per year. Studies reveal that roughly 15-30% of data center equipment consumes energy while idle. In fact, server utilization rates typically hover between a meager 5% and 15%, even though they continue to draw full power (Jahangir et al., 2021). This underscores a critical inefficiency, indicating that server utilization in data centers falls significantly short of optimal levels. To address the challenge of rising energy consumption, cloud computing leverages techniques like dynamic power management (DPM) and virtual machine (VM) consolidation. VM consolidation involves migrating workloads from underutilized systems to others, allowing idle servers to be powered off. This approach has demonstrably reduced peak power consumption of servers during idle states – from 50% to 20% over the past decade (Pierson and Hlavacs, 2015).

In addition, the rapid growth of the internet presents a significant challenge to achieving the goal of green computing. This rapid expansion, characterized by an

increase in the number of users, devices and data results in an unconventional source of increased energy consumption. According to a report by CISCO, global internet users are estimated to reach 5.3 billion by 2023, representing two-thirds of the world's population (Zhang et al., 2019). Furthermore, per capita device ownership is expected to reach 3.6 devices, with a total of 29.3 billion devices by 2023. Recent developments in the Internet of Things (IoT) are further accelerating the growth of the internet, with an estimated 14.7 billion smart devices, connected for communication, are expected to be operational by 2023 (Zhang et al., 2019). Similarly, the rapid growth in internet activity leads to sharp increase in data volume. For instance, approximately 300 million mobile applications have been downloaded by 2020 alone (Zhang et al., 2019). This exponential increase in data generation inflates the data volume that is projected to reach a staggering 150 zettabytes by 2024 (Kireev et al., 2019). On the other hand, in 2020, individual data generation reached an estimated 1.7 megabytes per second and 2.5 quintillion bytes per day. Underscoring the rapid growth, it is estimated that 90% of the world's data has been produced in just the last two years (Roser, 2022). These rising trends contribute to a growing data storage demand in the data centers (DCs), resultantly, storage alone accounts for 11% of total DC power consumption (Jahangir et al., 2021). Furthermore, the common practice of storing multiple copies of the same data within DCs significantly expands the data volume (Jahangir et al., 2021).

Therefore, the growing volume of big data and the challenge of data latency requires the use of well-established mechanisms such as data replication. In cloud environments, data replication plays an important role in achieving reliability and fault tolerance that ensures adherence to Service Level Agreements (SLAs). This process involves copying essential data closer to the client, minimizing the distance data must travel, and reducing latency. Data replication follows a three-phase process: staging, placing, and moving. However, a significant drawback of data replication is that once a specific node is activated, it cannot be deactivated, even when the node is idle. The key reason of the drawback lies in continuous operation stemming from the node's responsibility to maintain data availability causing a conflict with the conventional energy-saving techniques like VM consolidation and dynamic power management. Additionally, the increasing frequency of data replication results in a higher number of idle nodes hosting replicated data, leading to substantial energy wastage.

This study presents a novel approach that addresses the challenge of balancing data replication with energy efficiency in cloud computing. The proposed approach integrates two conflicting paradigms including energy efficiency and data replication. Energy efficiency requires shutting down underutilized nodes, whereas data replication aims to place replicated data on underloaded nodes for faster access (potentially saving time as compared to complex retrieval algorithms). Additionally, the study presents a mechanism that enables simultaneous operation of data replication and dynamic power management (DPM) including an intelligent data replication placement strategy. The placement strategy categorizes the nodes based on their current workloads and implements a tailored policy for each workload category. Based on CPU utilization, the workloads are categorized as underloaded, normally-loaded, and overloaded respectively. Underloaded nodes are powered off through DPM for energy efficiency, whereas the workload of overloaded nodes is balanced via a load balancer for optimal performance. However, neither underloaded nor overloaded nodes are considered while making decisions about the data replication placement. The data replication is hosted only upon the normally-loaded nodes that neither hinder the process of DPM nor

degrade performance during the data replication process by becoming unresponsive. Data replication is hosted only on the normally-loaded nodes, ensuring they neither impede DPM processes nor compromise the performance during the replication process due to unresponsiveness.

Furthermore, the decision of replication placement is dependent on factors such as CPU utilization, proximity to requesting clients, available bandwidth, and available memory. The proposed approach outlines the rationale for initial placement of the replica as well as continuously monitors the host for these factors even after the placement. In case of the current host become unsustainable, the replica is automatically migrated to a new, more suitable node. The proposed EnE-Rep introduces several key features for achieving balanced resource utilization and energy efficiency in cloud data center given as following:

- Introduction of a framework that categorizes hosts within the cloud data centers into underloaded, normally-loaded, and overloaded based on the workload.
- Implementation of a double threshold mechanism that activates the load manager and energy monitor in response to the dynamic and unpredictable workloads typically encountered in cloud data centers.
- Integration of a replicator module capable of intelligently selecting an energy-efficient node for replica placement based on the factors such as CPU utilization, proximity, bandwidth, and memory.
- Development of an architecture incorporating VM selection methods for facilitating VM migration from overloaded and underloaded hosts.
- Evaluation of the proposed algorithm's performance using CloudSim, and Planetlab (a real-world workload consisting of 800 cloud data centers distributed across 500 distinct locations worldwide).
- Comparative analysis of the results with an approach that employs intelligent placement of data replication based on popularity for energy consumption.

Section 2 presents the related work; Section 3 clearly defines the problem statement; Section 4 introduces the proposed EnE-Rep model; Section 5 describes the experimental setup used for evaluation; Section 6 presents results and relevant discussion; and finally, Section 7 presents the conclusion based on the discussion in section 6 and potential avenues for future work.

2. Related Work

Data replication involves the decisions regarding the creation, storage, placement, and processing of a necessary replica. Replication decisions, which vary based on context including centralized, distributed, offline, or online, significantly affect the system performance and user experiences. Similarly, replication placement is an important aspect of data replication, particularly, the decision regarding the optimal location for transfer the replica poses a significant challenge. Therefore, placement scheduling should carefully be managed for preventing network congestion, ensuring replica availability, and maintaining efficient access times. A concise overview of the relevant studies is presented as following:

Atrey et al. (Atrey et al., 2019) proposed a scalable placement strategy for distributed cloud storage systems which partitions the data to manage large workloads efficiently. The researchers incorporate two scalable algorithms for efficiently addressing the computational demands. By partitioning data, the revised model enhances the system scalability and resource utilization. In addition, the model enhances system performance by reduces processing time and computational cost. However, the effectiveness of the partitioning model may vary depending upon data characteristics which necessitates the maintenance of data integrity and accessibility. Additionally, the algorithms may introduce complexity and potential trade-offs in terms of accuracy and resource usage. Similarly, Zhang (Zhang, 2020) introduced a time-efficient multi-objective approach for the replication placement problem in cloud storage systems by prioritizing Quality of Service (QoS) restrictions to minimize system response time. The proposed approach ensures an improved user experience and meets performance requirements by implementing QoS restrictions, as well as providing a balanced solution through the simultaneous optimization of various factors. However, potential drawbacks of the proposed approach include the complexity of the optimization process, challenges in meeting all QoS restrictions, and the assumption that minimizing the response time is always the primary objective, which may not align with other system requirements or trade-offs.

Subsequently, Ao and Psounis (2020) proposed a framework for efficient resource allocation in cloud computing systems for handling hierarchical and heterogeneous tasks. The framework minimizes task completion time by leveraging two key strategies including data replication for system reliability and a hierarchical resource management structure for optimizing performance. However, the framework's effectiveness depends on precise resource allocation algorithms and workload characterization. Inaccurate or inefficient allocation methods may lead to suboptimal task completion times. In addition, the hierarchical structure may introduce additional complexity and overhead. On the other hand, Huang et al. (Huang et al., 2020) proposed a mining-based approach for discovering interactions between data entities in cloud storage. The proposed approach aims to improve efficiency and reduce energy consumption by optimizing resource allocation. Additionally, the mining approach incorporates replica placement and backup for enhanced data availability and fault tolerance. However, inaccurate capture of interaction and relevant relationships may limit the potential efficiency gains. Moreover, replica placement and backup require additional storage space and computing resources.

Next, Bacis et al. (Bacis et al., 2019) proposed a data management approach for cloud storage that guarantees data availability and confidentiality during node failures. The proposed approach leverages "all-or-nothing" transformations and fountain codes. All-or-nothing transformation secures data integrity and confidentiality through encryption, whereas fountain codes enable data recovery from transmission errors or failures. However, weak encryption or inefficient fountain codes may compromise security or data availability in addition to the computational overhead by encryption and decoding processes. Similarly, Khalili Azimi (2019) proposed a data management approach based on a bee colony optimization for enhancing data availability in cloud storage. The bee colony optimization algorithm provides a decentralized and self-organizing approach, mimicking the behavior of a bee colony to efficiently search for optimal replication configurations. Consequently, the system demonstrates robustness against the changing conditions and optimize replica placement based on factors such as data importance, workload patterns, and resource availability. However, accurate

decision-making is significant, as inefficient choices can result in wasted resources. Moreover, Edwin et al. (Edwin et al., 2019) introduced a dynamic and cost-effective data replication approach that enhances data availability and the replication process. The data replication approach utilizes a multi-objective optimization scheme that prioritizes cost-effective replication by considering replica costs in various data centers. In addition, the knapsack algorithm is enhanced to balance availability and load during replication, optimizing cost-effectiveness and load balancing. By dynamically adjusting replication levels based on cost and availability, the proposed approach optimizes resource utilization and reduces unnecessary overhead. However, performance of the data replication approach depends on the accuracy of the cost model and the knapsack algorithm; inaccurate cost estimates may result in suboptimal replication decisions, thereby impacting cost-effectiveness. Furthermore, Mostafa (2020) introduced a data replication consistency method for cloud-fog environments for improving system availability, fault tolerance, and Quality of Service (QoS). The research aims to prioritize the preparation of the system for potential availability issues to ensure continuous service. The implementation of data replication consistency enhances fault tolerance, minimizing data loss and disruptions, which leads to a more reliable and consistent user experience (QoS). However, inadequate or inconsistent replication can cause data inconsistencies. Additionally, the trade-off between system availability and resource utilization should be carefully managed to avoid excessive replication overhead.

On the other hand, Ramanan and Vivekanandan (2019) investigated the security vulnerabilities in cloud systems using a stochastic diffusion search algorithm for optimizing data replication costs. The stochastic algorithm promotes efficient resource utilization and cost savings by intelligently distributing replicas based on dynamic factors such as workload, resource availability, and network conditions. In addition, the stochastic diffusion algorithm strengthens cloud system security, safeguarding sensitive data from unauthorized access. However, aggressive cost reduction through inaccurate modeling or the algorithm's inherent randomness (stochastic nature) may pose scalability challenges in large cloud deployments. Conversely, Abbes et al. (Abbes et al., 2020) explored virtualizing container concepts for distributed applications in cloud storage. The research predicts replication factors (i.e., number of copies) needed for maintaining availability during container failures using experimental forecasting based on regression analysis. Although, virtualization improves resource utilization and scalability for containers, however, the regression approach used for replica placement relies heavily on the quality of data, assumed linear relationships, potentially overlooking various factors affecting availability.

Alternatively, Tahir et al. (Tahir et al., 2021) addresses user privacy and data integrity concerns in cloud systems using a Genetic Algorithm (GA) for generating encryption and decryption keys. The proposed tailored approach enhances data security and user privacy, ensuring the confidentiality of sensitive information. However, the computational complexity of the GA approach may strain system resources, potentially affecting performance. Furthermore, safeguarding the secure storage and management of generated keys is essential for upholding data integrity and privacy. Subsequently, Babar et al. (Nazir et al., 2018) proposed the CDSS-RPS data replication system, a two-phase approach for optimizing replica placement and file access time in cloud storage. In first phase, a centralized decision system, leverages node computing capacity for optimal replication placement. On the other hand, the second phase considers factors like access frequency, storage capacity, and response time for improving access time. Although,

Gridsim-based implementation validates the effectiveness of the two-phase CDSS-RPS data replication system, however, accurate estimations of computing capacity and response time are significant in managing replica placement and balancing file access. Finally, Ebadi et al. (Tagne Fute et al., 2023) proposed a hybrid heuristic called, Hybrid Particle Swarm Optimization Tabu Search (HPSOTS) for intelligent data replica placement. Due to the trade-off between replication and energy efficiency, the proposed research categorizes the problem as NP-hard. HPSOTS is a nature-inspired algorithm that achieves significant improvements in Total Energy Consumption (TEC) and cost as compared to existing approaches. By evaluating multiple options for fulfilling read or write requests based on energy consumption, the study lays the foundation for our proposed work. A brief summary of most related studies is presented in Table I.

3. Problem Statement

Database replication is an important approach in cloud computing for improving data access times. However, in dynamic and heterogeneous cloud environments with unpredictable workloads, existing data replication scheduling approaches can lead to inefficiencies:
Overloaded Hosts: Replication tasks scheduled on overloaded hosts can increase data access times due to the computational overhead of complex replication algorithms, potentially violating Service Level Agreements (SLAs).
Underloaded Hosts: Replication tasks placed on underloaded hosts lead to wasted energy consumption. These hosts cannot be powered down for energy savings due to the ongoing replication tasks they support for other nodes. This combined effect leads to increased energy consumption and potential SLA violations in cloud data centers.

3.1. Research Objectives

This study aims to develop a novel data replication scheduling approach that addresses the limitations of existing methods by:
Optimizing Resource Allocation: The proposed approach seeks to consider real-time workload information for scheduling replication tasks on suitable hosts, avoiding overloaded nodes.
Minimizing Energy Consumption: Replication tasks are intended to be placed on underloaded hosts that are likely to be powered down for energy savings. By addressing these challenges, the proposed approach can lead to significant reductions in energy consumption and improve the overall efficiency of data replication in dynamic cloud environments.

Table 1. Comparative Summary of Related Work

Article	Type of Handling	Place of Handling	Performance Metrics	Evaluation Tools	Limitations
(Atrey et al,. 2019)	Placement	Server	Delay	CPP, Python	Weak presentation
(Zhang, 2020)	Placement	Server	Response time QoS	Implementation	Weak and inconclusive model implemented in unfamiliar environment.
(Ao and Psounis, 2020)	Modeling	Server	Cost Delay	Analytical	Highly complex optimization problem.
(Huang et al., 2020)	Placement	Multiple	Energy	MapReduce	Unrealistic system model
(Bacis et al., 2019)	Management	3rd party	Availability Security	Storj	Unclear model weak organization
(khalili azimi, 2019)	Management	Server	Throughput Delay	MATLAB	Unclear model weak organization
(Edwin et al., 2019)	Management	3rd party	Cost Availability Energy	Cloudsim	Increases the overall overhead when increasing replicas.
(Ramanan and Vivekanandan, 2019)	Security	Server	Security Cost QoS	Simulation	Shortage in experimental results
(Abbes et al.; 2020)	Modeling	3rd Party	Availability	Implementation	Malfunctioning of regression around 0 values of prediction.
(Tahir et al., 2021)	Security	Client	Security Delay	Implementation	Shortage in experimental results
(Fan et al., 2021)	Consistency	Multiple	Response time	Cloudsim	Not suitable for real-time data

4. Research Methodology of the Proposed EnE-Rep Model

This study introduces an inclusive model designed to optimize the benefits of the data replication process while simultaneously reducing the overall system energy consumption. The model, called EnE-Rep, categorizes nodes into three distinct groups: underloaded, normally-loaded, and overloaded. For each category, a tailored strategy is implemented such that overloaded nodes are balanced, whereas underloaded nodes are efficiently shut down using virtual machine consolidation and dynamic power management methods for energy efficiency. Similarly, this section provides a comprehensive overview of the components and nuances comprising the architecture of the proposed model. Major subsections present the discussion on topics such as replication request submission, SLA checking, utilization management, load management, energy management, sorting management, and periodic recursion monitoring respectively.

Nevertheless, the underlying system prioritizes honoring Service Level Agreements (SLAs) for time-constrained users. An initial check ensures any optimization process won't introduce delays that could violate these SLAs. Subsequently, the method leverages a heuristic-based approach to determine CPU utilization for each node (Hastie et al. 2009). The heuristic-based approach is then utilized for categorizing the nodes as overloaded, underloaded, or normally-loaded. CPU utilization exceeds 85% in the overloaded nodes, whereas it remains below 30% in the underloaded nodes (Beloglazov et al., 2012; Hastie et al., 2009).

Once the workload of a node is determined, the overloaded nodes are directed to the load balancer module. The load balancer module identifies the least occupied node from the entire host list and creates a new VM on a suitable node for transfering the excessive load. Similarly, the workload from underloaded nodes is migrated, and the nodes are vacated. These vacated nodes are subsequently powered off to reduce the overall energy consumption of the system. In the final phase of the proposed approach, normal nodes are sorted in ascending order based on their CPU utilization, proximity, bandwidth, and available memory. Percentile values from all sorted lists are standardized to bring them onto the same scale. The weighted average, calculated as the summation of the product of weights and quantities divided by the summation of weights, is determined according to Eq. (1).

$$\boldsymbol{Weighted\ Average} = \frac{\sum(\boldsymbol{Weights}\times\boldsymbol{Quantities})}{\sum \boldsymbol{Weights}} = \frac{\sum_{i=1}^{n} \boldsymbol{w_i x_i}}{\sum_{i=1}^{n} \boldsymbol{w_i}} \quad (1)$$

where $\boldsymbol{w_i}$ represents the weight of the objective in a priority-based arranged list and $\boldsymbol{n}$ indicates the total number of objectives. Subsequently, leveraging the weighted average calculations from Eq. (1), EnE-Rep creates a weighted average list for all normal nodes in contention for replica placement using Eq. (2).

$$\boldsymbol{W_{Avg}} = 40 * CPU + 30 * Prox + 20 * BW + 10 * \quad (2)$$

where weights ranging from 40 to 10 are assigned to factors influencing replica placement, with higher weights indicating greater influence on the final score. Similarly, $\boldsymbol{CPU}$ represents the CPU utilization of the node, reflecting the node's processing capacity. On the other hand, $\boldsymbol{Prox}$ shows the proximity of the node to the requester(s) who will access the replica, whereas the bandwidth $\boldsymbol{BW}$ represents data transfer capabilities between the node and requesters. Finally, available memory ($\boldsymbol{RAM}$) on the node is important for storing replica data effectively. The node with the lowest score on this list is selected to host the data replica. This selection approach prioritizes a balance between resource utilization, data access speed, and energy efficiency. A detailed explanation of each sub-module within EnE-Rep is provided in the following sections.

4.1. Replication Request Submission (RRS)

The RRS module initiates the data replication process under two primary conditions. Firstly, any modification made to the original data (denoted as "t") triggers replication, ensuring all replicas are updated with the latest version. Secondly, when a remote user

accesses data from a remote replica repository and requests an updated copy, replication is triggered to provide the user with the most recent version of the data.

4.2. SLA Manager

The SLA Manager prioritizes adherence to Service Level Agreements (SLAs) for time-constrained users. Given the complex algorithms involved in data replication to ensure proximity to the requester, the SLA Manager identifies and separates cloudlets with time constraints from those operating under more flexible timeframes. Resultantly, the exclusion of time-constrained nodes from subsequent optimization steps effectively prevents potential SLA violations, thereby ensuring the fulfillment of their SLAs.

4.3. Utilization Manager

The utilization manager in EnE-Rep plays an important role by conducting a comprehensive analysis of CPU utilization across all nodes prior to scheduling data replication. This analysis serves as a critical filtering mechanism where overloaded nodes surpassing a utilization threshold are routed to the load balancer module for resource optimization, and underloaded nodes with low utilization are earmarked for potential migration and energy conservation through the energy monitor module. Finally, nodes with balanced CPU utilization are directed to the replicator section for data replication tasks. This intelligent allocation process ensures optimal resource utilization and prevents overloading nodes with replication tasks.

4.4. Load Monitor

The Load Monitor, receiving a list of overloaded nodes from the Utilization Manager, acts as a pivotal task reassignment unit for ensuring workload distribution across the system and prevent resource bottlenecks. The operations of Load Monitor consist of three main steps; first, it identifies suitable underloaded hosts from the entire host pool, considering factors like available CPU capacity, memory, and bandwidth. Secondly, Load Monitor assesses the projected workload on the candidate host post-load transfer, ensuring it remains below a predefined upper threshold to prevent overloading. Upon passing the feasibility check, the Load Monitor executes the workload transfer, potentially involving the creation of a new virtual machine (VM) on the underloaded host. Finally, after completing load balancing via VM consolidation and Dynamic Power Management (DPM), the Load Monitor forwards an updated list of "normalized" hosts—those with balanced workloads—to the Replicator module for optimal replica placement decisions.

4.5. Replicator

The replicator module serves as the focal point for determining data replication placement, considering four key factors: CPU utilization, proximity, bandwidth, and available memory across all hosts. To facilitate fair comparison, lists corresponding to each property are created and processed through a percentile calculator by aligning units across diverse metrics. Subsequently, weights for each property are computed using Eq.

(2), and their weighted average is calculated. This process identifies the optimal host for data replication, ensuring efficient resource utilization. Furthermore, the hosting VM's priority is elevated to mitigate any delays incurred during decision-making that guarantees swift data access.

4.6. Energy Monitor

The proposed model comprises several key components, each playing a significant role in optimizing system efficiency. Firstly, underloaded nodes identified by the utilization manager undergo dynamic power management, where idle nodes are shut down to decrease overall energy consumption. This process involves transferring the workload of nodes below the CPU utilization threshold to other suitable hosts based on CPU, bandwidth, and memory considerations before shutting them down, as depicted in Algorithm 1. Additionally, Figure 1 illustrates the comprehensive architecture of the model, depicting its major components and their interactions. Figure shows the end user's interaction with remote hosts where cloud computing services are accessed. Behind these remote hosts, the proposed methodology's operational intricacies are implemented. Following optimization, the relevant data is integrated into the central database.

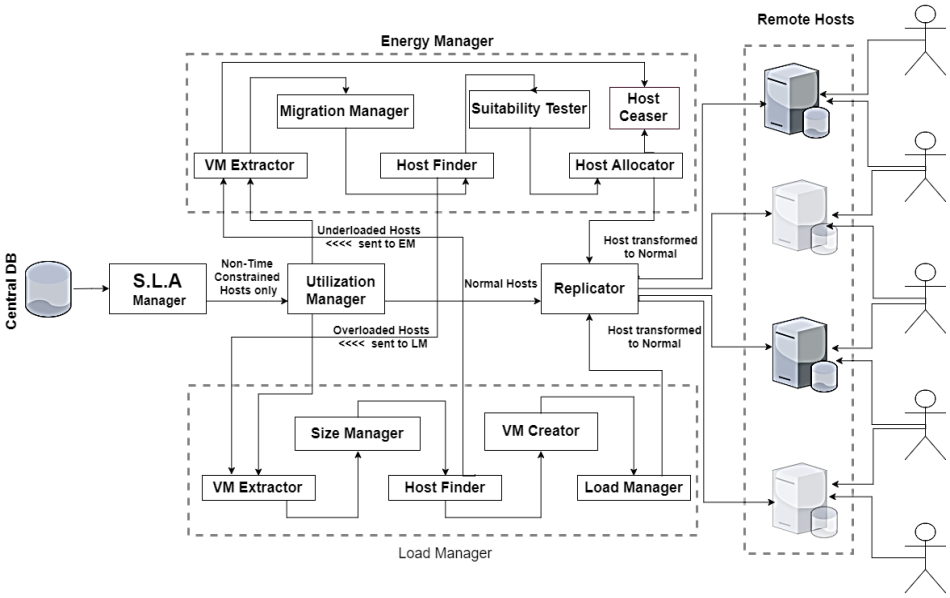


Figure 1. Detailed Architecture and Interaction Diagram of EnE-Rep

The algorithm for energy-efficient data replication outlines the optimization mechanism applied to non-time constrained cloudlets. CPU utilization is prioritized, with heavily utilized nodes given precedence. Lists for proximity, bandwidth, and memory are sorted and transformed into percentile lists to standardize units. The weighted average formula generates an energy-efficient list, with component weights determined by their perceived importance.

Algorithm 1: EnE-Replication for Normal Ranged Nodes

Input: Hosts within the utilization range <Normal >
Output: Replica placement on the BEST among <Normal >hosts

```
1                                                    cloudlet.makeReplica()
2  forall hostsinhostList do
3    if replica==True then
4      if TC==False then
5        cputilSorted=getUtil( hostList)
6        Utilization list of hosts is created and sorted ascendingly
         As greater is preferred
7        proxSorD=getProx(hostList)
8        Proximity list of all hosts is created and sorted descendingly
         As lesser is
9        bwSor=getBw (hostList )
10       Bandwidth utilization list is created and sorted ascendingly
         asgreaterispreferred
11       ramSor=getRam(hostList)
12       RAM utilization list of all hosts is created and sorted
         Ascendingly as greater is preferred
13
14       cputilPercen← calPercen (cputilSorted)
          proxPercen←calPercen (proxSorD)
          bwPercen←calPercen (bwSor)
          ramPercen←calPercen(ramSor)
15
16       eeList.add(i)=40*cputilPercen.get(i)+30*proxPercen.get(i)
          +20*bwPercen.get(i)+10*ramPercen.get(i);
17       forall itemineeList do
18         if current < previous then
19           Best← current
                          // smallest element is searched
20         end
21         Best← previous
22       end
23       allocateReplica(Best.getId())
            // replica is placed on Host which is BEST w.r.t
          all four (Util.Proximity.BW.RAM)parameters
24       setPriorityHigh(getReplicatedVm())
           // Priority of VM dealing replica on BEST host is
          set to High so that scheduler gives it PE early
          and Max available to counter any data access delay
25     end
26   end
27 end
```

The algorithm then ranks normally-loaded nodes based on their total score across all parameters, selecting the most suitable host for data replication. This approach aims to maximize energy savings by efficiently utilizing system resources and minimizing idle states. To compensate for optimization time, the host node's priority is set to high.

The algorithm for energy-efficient data replication in data-intensive clouds outlines the operational framework of the proposed approach. Upon receiving a cloudlet with a data replication scheduling request, the model employs its optimization mechanism that is tailored for energy-efficient placement through exclusive attention to non-time constrained cloudlets. Initially, the algorithm retrieves CPU utilization data and arranges it in descending order to prioritize highly occupied nodes. Similarly, sorted lists are generated for proximity, bandwidth, and memory along with the percentile lists are established for standardizing the units. Nodes with the highest numerical values for each parameter top their respective lists, and an energy-efficient list is computed using a weighted average formula. Component weights are assigned based on perceived significance; however, CPU utilization is prioritized due to its relevance to workload segregation. The algorithm evaluates all four weighted average values for each node to rank them based on their collective scores. Among normally-loaded nodes, those with the highest scores are deemed optimal for data replication placement, striking a balance between workload and energy efficiency. This strategy facilitates the idling and shutdown of underloaded nodes, contributing to significant energy savings. Finally, host node priority is elevated for optimization time that ensures efficient scheduling of data replication tasks.

5. Experimental Setup

The Infrastructure as a Service (IaaS) model in cloud computing offers extensive computing resources with advantages like repeatability and resource control which necessitates thorough testing of proposed data replication approach on large-scale Data Centers (DCs). However, physical platforms of such magnitude are challenging to procure, prompting the use of simulation. Leveraging Cloudsim toolkit v3.0 proves ideal for this purpose that is tailored for cloud environments and sparing users from intricate details. Cloudsim facilitates dynamic workload integration through the inclusion of energy consumption modeling and accounting functionalities. Following is a detailed overview of the infrastructure setup and the submitted jobs for simulation:

5.1. Resource modeling

This study utilizes the CloudSim Toolkit 2.0 platform (Beloglazov *et al.* 2012), developed by Beloglazov and Buyya, for simulating a data center (DC) environment. The simulated DC comprises 800 Physical Machines (PMs), with half being HP ProLiant ML110 G4 servers and the other half HP ProLiant ML110 G5 servers. Table 2 provides detailed specifications regarding RAM and Processing Element (PE) for these server types. The server models such as HP ProLiant ML110 G4 and G5, demonstrate varying RAM and PE specifications, as presented in Table 2. Similarly, the processing power, measured in MIPS (Million Instructions Per Second), varies between the server models. The HP ProLiant ML110 G4 delivers 1860 MIPS, whereas the G5 model is more powerful at 2660 MIPS. Additionally, each server is allocated a bandwidth of 1000

MBs. Virtual Machines (VMs) in this experiment emulate Amazon EC2 instances, however, configured with a single core. The simulations are conducted on actual hardware platforms, comprising HP ProLiant and IBM GX3250 machines.

Table 2. Resource Specification of the Servers used in Simulation

Instance Type	Specification
Extra Large	2000 MIPs, 3750 MB
Medium	2500 MIPs, 850 MB
Small	1000 MIPs, 1700 MB
Micro	500 MIPs, 613 MB

5.2. Application modeling

Table 3 shows the parameters adjusted to create distinct workload scenarios, presented in ascending order of intensity. The application model utilizes authentic data sourced from the PlanetLab project that is specifically gleaned from traces of over 1000 VMs allocated to diverse users. These traces, derived from PlanetLab's CoMon Project spanning 10 days, depict authentic workload patterns. The rationale behind employing linear workload variations stems from the understanding that power consumption correlates linearly with factors such as CPU utilization, memory usage, storage access, and network activity. This methodology enables the evaluation of the EnE-Rep model's scalability across varying workload intensities.

5.3. Performance evaluation parameters

EnE-Rep differs from traditional replication considerations by placing a primary emphasis on energy efficiency over factors such as cost, response time, and reliability. On the other hand, conventional approaches prioritize various performance metrics as compared to EnE-Rep's novel strategy revolves around minimizing power consumption. Studies identify the direct correlation between a system's power usage and factors including CPU utilization and memory usage (Beloglazov et al. 2012; Fan et al., 2007; Kusic et al., 2009). In addition, data access time is influenced by factors like distance and available bandwidth. However, EnE-Rep introduces a unique energy conservation method by strategically migrating virtual machines (VMs) from specific hosts, enabling their shutdown to conserve energy. Subsequently, to assess the energy-saving benefits, EnE-Rep evaluates key metrics including the total number of VM migrations, successful host shutdowns, and components of data access time such as VM selection, host selection, and VM relocation time. EnE-Rep actively evaluates its energy-saving effectiveness through several key metrics. These metrics include the number of VM migrations enabling host shutdowns, and the various components of data access time including VM selection, host selection, and VM relocation time.

Table 3. Workload Variations Applied in the Experiment

Workload Set	Job file size (Bytes)	Job length (MI)	No. of VMs	No. of Hosts
1	300	2500	1000	1000
2	650	5000	2000	2000
3	1000	7500	3000	3000
4	1300	10000	4000	4000
5	1500	13000	5000	5000
6	1800	16000	6000	6000
7	2200	20000	7000	7000
8	2600	25000	8000	8000
9	3000	30000	10520	8000

5.4. Energy consumption

Cloud data centers are major consumers of energy that is primarily attributed to CPUs, storage disks, and network equipment, with CPUs being the most power-intensive components. Traditionally, techniques like Dynamic Voltage and Frequency Scaling (DVFS) have been employed to mitigate CPU power consumption through adjusting operating frequency and voltage. Despite its near-linear relationship between power and frequency, DVFS is limited by the finite number of available frequency states. In contrast, EnE-Rep adopts a more significant approach by powering down idle nodes based on the notion that around 70% of power is consumed by idle resources. Leveraging this strategy enables EnE-Rep to achieve greater energy savings compared to DVFS. Energy consumption is measured using Eq (3) (Cidon *et al.* 2013) given as following:

$$P(u) = K * P_{max} + (1 - K) * P_{max} * U \tag{3}$$

where P_{max} denotes the maximum power consumption under full server utilization, K represents the fraction of power consumed by the idle server, and U signifies the CPU utilization. Subsequently, energy is computed using Equation (4) (Bagheri and Mohsenzadeh, 2016) given as following:

$$E = \int_0^{\infty} P(u)\ dt \tag{4}$$

Where E represents the total energy consumption over the time period starting from t and extending indefinitely into the future. Similarly, $P(u)$ denotes the power consumption, which is a function of the CPU utilization $u(t)$. The function $P(u)$ gives the power consumed by the system at any given time t. Subsequently, t is the lower limit of the integral, representing the starting time from which the process of measuring

energy consumption has begun. Finally, ∞ highlights the upper limit of the integral, indicating that the energy consumption is being considered over an infinite time period, essentially summing up the power consumption from time t to the end of time (or theoretically, forever).

Validation of the Equation

It is important to consider the context in which equation (4) is applied. The equation assumes that the system, such as a server, operates continuously starting from time t without a defined endpoint. This assumption is particularly relevant for systems like cloud servers, which are often designed to run indefinitely. Additionally, the power consumption $P(u)$ is time-dependent because the CPU utilization $u(t)$ varies over time. Equation (4) accounts for this variability, recognizing that power consumption is not constant but fluctuates with the level of CPU usage at any given moment. Furthermore, the integral in the equation accumulates the total energy consumed over the period from time t to ∞. Since energy is the product of power and time, integrating the power over this period yields the total energy consumption, providing a comprehensive measure of the system's energy usage.

The choice of ∞ as the upper limit in the integral can be justified on several grounds. Firstly, it ensures theoretical completeness by covering the entire potential lifespan of the system, thus accounting for all possible future energy consumption. This is particularly relevant in theoretical models where the system is assumed to operate indefinitely. Secondly, using ∞ as the upper limit is essential for modeling long-term energy consumption, especially in systems like cloud data centers which are designed for continuous operation. This employed approach aids in understanding long-term energy consumption patterns, which is important for making informed decisions about energy efficiency, sustainability, and cost management. Additionally, integrating up to ∞ enables worst-case scenario analysis by estimating the maximum possible energy consumption over time, which is valuable for planning purposes such as provisioning energy resources and designing cooling systems.

For comparison, the energy consumption is calculated as

$$EnergyConsumption\left(\frac{Kw}{h}\right) = \frac{EnergyConsumption}{3600 * 1000} \tag{5}$$

6. Results and discussion

This section presents a detailed performance evaluation of the proposed EnE-Rep model against the classical scheduling policies and a metaheuristic technique called Hybrid Particle Swarm Optimization Tabu Search (HPSOTS). Figure 2 presents a heatmap that visually compares the number of VM migrations required by different scheduling techniques for all seven tested scheduling policies. Figure shows that the scheduling policies without optimization (thrrs, iqrmmt, iqrrs, lrrs, madmc, thrmc, and thrmu) suffer from significantly higher VM migrations, as indicated by the darker shades in the heatmap. This is due to their less effective approach of placing replications on the first available host without considering the load or suitability of the host. However, HPSOTS exhibits a reduction in VM migrations as compared to non-optimized techniques.

HPSOTS applies some level of intelligence in selecting hosts for replication placement, potentially considering factors that contribute to energy consumption. On the other hand, EnE-Rep demonstrates the most significant reduction in VM migrations as compared to both non-optimized scheduling methods and the HPSOTS metaheuristic technique. This prominent improvement is evidenced by the decrease in migrations from a staggering 44200 to a more manageable 25349. The success of EnE-Rep in minimizing migrations can be attributed to its intelligent approach to replica placement. EnE-Rep adopts a double threshold policy for CPU utilization, ensuring that replications are only placed on hosts with CPU usage within a specific, optimal range. By avoiding overloaded hosts, EnE-Rep eliminates the need for frequent migrations that is caused by the performance bottlenecks which results from insufficient resources. Additionally, by steering clear of underloaded hosts, EnE-Rep prevents unnecessary migrations triggered by inefficient resource allocation on underutilized machines.

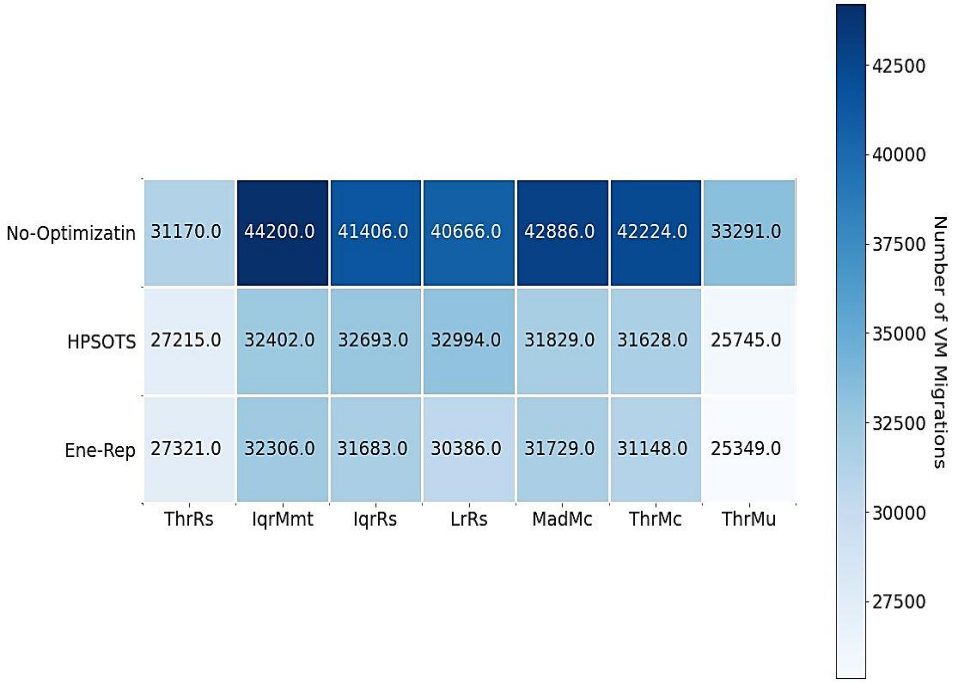


Figure 2. Heatmap for number of VM migrations during the execution

Similarly, Figure 3 explores another important aspect of VM migrations – the mean time before a VM migration becomes necessary. The analysis in Figure 3 compares how long VMs stay on a host before needing to be migrated. Unsurprisingly, non-optimized policies perform inefficntly due to their lack of migration consideration. The high frequency of unnecessary migrations in these policies directly affects their performance. However, HPSOTS prioritizes energy efficiency by evaluating the entire host population, nonetheless, it might not prioritize factors that directly reduce the number of

VM migrations. On the other hand, EnE-Rep shows better performance as compared to classical scheduling policies, particularly from HPSOTS by its adept optimization of VM migration frequency. The optimization is accomplished through a targeted approach: firstly, by assessing the CPU utilization of potential host candidates, and secondly, by prioritizing the placement of replications exclusively on hosts with normal CPU loads. This meticulous methodology yields several notable advantages. Firstly, it leads to reduced disruptions by allowing VMs to remain on suitable hosts for extended durations, thus mitigating the need for frequent migrations. Secondly, it enhances system performance by avoiding overloaded hosts, thereby averting potential performance degradation that is caused by the resource bottlenecks which culminates from frequent migrations.

Subsequently, Figure 4 presents a comparison of the time taken by each method to select a suitable host for data replication placement. HPSOTS exhibits the longest selection time because it evaluates the entire host population and ranks them based on energy consumption, thereby prioritizing comprehensive analysis over speed. In contrast, both EnE-Rep and the non-optimized methods demonstrate relatively similar selection times.

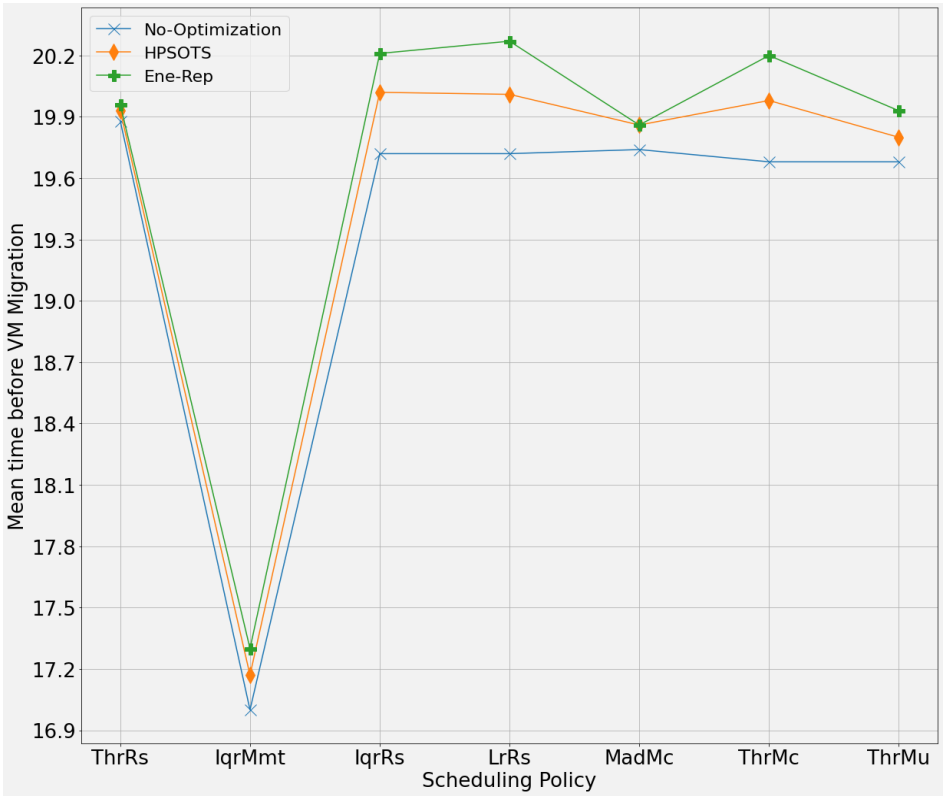


Figure 3. Mean time before a VM migration throughout the execution

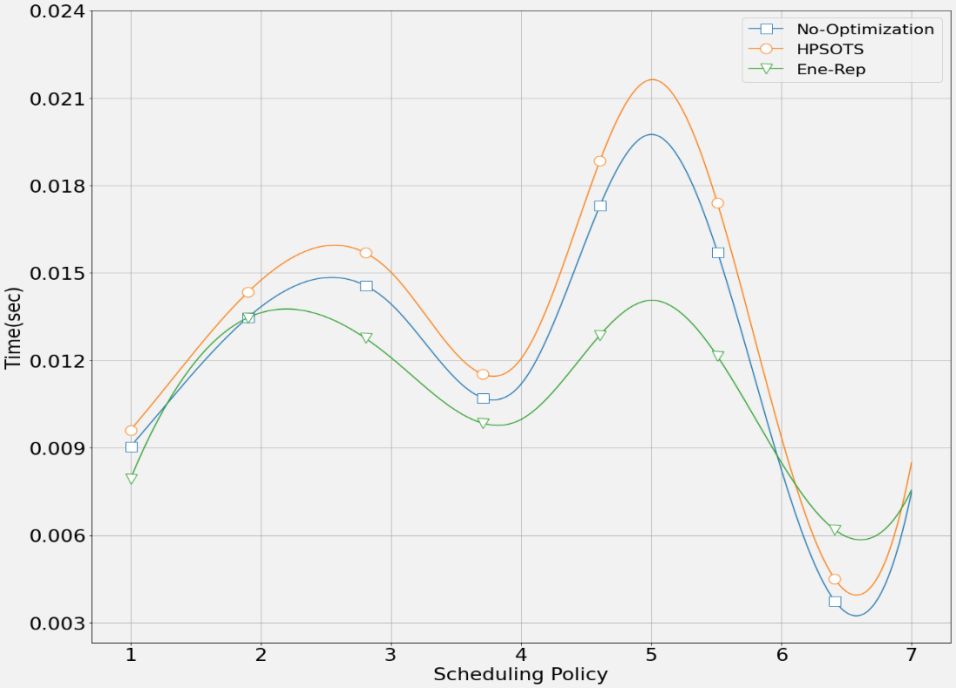


Figure 4. Mean time for host selection for placement of data replication

These approaches share a key characteristic, i.e., these methods focus on evaluating a single candidate host at a time, rather than the entire pool. Therefore, once a suitable host is found and meets the requirements, the replication is placed, and the selection process ends. However, in non-optimized techniques, if the initial candidate is unsuitable, an iterative search might be necessary to find an alternative that leads to longer selection times. EnE-Rep, on the other hand, leverages a predefined CPU utilization threshold, allowing it to identify suitable hosts faster as compared to the non-optimized technique's potentially time-consuming iterative search. By focusing on a specific CPU utilization range, EnE-Rep efficiently narrows down potential candidates that results in shorter selection time.

Finally, Figure 5 illustrates the energy consumption patterns of the proposed EnE-Rep against the other methods across all seven scheduling policies, providing a comprehensive view of their energy usage. The non-optimized techniques, represented by the blue bars, exhibit notably higher energy consumption in kilowatts. This heightened consumption can be attributed to two main factors. Firstly, non-optimized techniques trigger a substantial number of VM migrations, resulting in significant performance degradation. Additionally, these techniques adopt an unintelligent approach to replication placement, indiscriminately utilizing any available host regardless of its current workload. The utilized indiscriminate placement leads to disadvantages such as the replications placed on overloaded hosts trigger frequent migrations to address performance bottlenecks caused by insufficient resources. Conversely, placing

replications on underloaded hosts results in wasted energy consumption because these machines remain powered-on despite having minimal workload. Resultantly, overall increase in energy consumption is observed in non-optimized techniques. In contrast, while HPSOTS focuses on energy reduction, it does not explicitly consider the impact of replication placement on migration frequency. This oversight may result in the selection of energy-efficient hosts that are not optimal in terms of migration that can limit HPSOTS's overall energy savings as compared to EnE-Rep.

Figure 5. Energy consumption comparison of different strategies

7. Conclusion and future work

Cloud computing offers several advantages such as ease of use, affordability, adaptability, growth potential, and dependability, however, its growing infrastructure demands more energy and raises network distribution challenges. In this study, we have proposed EnE-Rep that integrates dynamic power management (DPM) and data replication for optimizing energy usage and enhance performance in simulations.

The performance evaluation of the EnE-Rep model against classical scheduling policies and the HPSOTS metaheuristic technique highlights its effectiveness in minimizing VM migrations and reducing energy consumption in cloud computing environments. EnE-Rep's intelligent replica placement strategy, guided by a double threshold policy for CPU utilization, effectively avoids overloaded and underloaded hosts, thereby mitigating the need for frequent migrations caused by performance

bottlenecks. In contrast, non-optimized techniques exhibit higher VM migration frequencies and performance degradation. Although, HPSOTS prioritizes energy efficiency, however, its oversight of migration reduction may limit its overall energy savings compared to EnE-Rep. Additionally, EnE-Rep's efficient host selection process results in shorter selection times as compared to non-optimized techniques. The analysis highlights the drawbacks of non-optimized approaches and emphasizing the importance of intelligent replica placement in reducing energy consumption. Furthermore, the fusion of data replication and energy efficiency in EnE-Rep presents promising avenues for greener and more stable ICT infrastructures.

Future work could explore proactive threshold strategies and decentralized approaches to enhance performance in stochastic cloud computing environments, ultimately advancing the goal of sustainable and efficient technology infrastructure.

Abbreviations

DC	Data center
DPM	Dynamic power management
DVFS	Dynamic Voltage and Frequency Scaling
GA	Genetic Algorithm
HPSOTS	Hybrid Particle Swarm Optimization Tabu Search
MIPS	Million Instructions Per Second
PE	Processing Element
PMs	Physical Machines
QoS	Quality of Service
SLAs	Service Level Agreements
TEC	Total Energy Consumption
VMC	Virtual machine consolidation

Acknowledgments

The author extends his appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through general project under grant number GRP/51/44

References

Abbes, H., Louati, T., Cérin, C. (2020). Dynamic replication factor model for Linux containers-based cloud systems. *Journal of Supercomputing*, **76**(9), 7219–7241.

Ao, W. C., Psounis, K. (2020). Resource-Constrained Replication Strategies for Hierarchical and Heterogeneous Tasks. *IEEE Transactions on Parallel and Distributed Systems*, **31**(4), 793–804.

Atrey, A., Van Seghbroeck, G., Mora, H., De Turck, F., Volckaert, B. (2019). SpeCH: A scalable framework for data placement of data-intensive services in geo-distributed clouds. *Journal of Network and Computer Applications*, **142**, 1–14.

Bacis, E., De Capitani DI Vimercati, S., Foresti, S., Paraboschi, S., Rosa, M., Samarati, P. (2019). Dynamic allocation for resource protection in decentralized cloud storage. In *2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings*, IEEE, , pp. 1–6.

Bagheri, K., Mohsenzadeh, M. (2016). E2DR : Energy Efficient Data Replication in Data. *Journal of Advances in Computer Engineering and Technology*, **2**(3), 27–34,,.

Balakrishnan, S. R., Veeramani, S., Leong, J. A., Murray, I., Sidhu, A. S. (2017). High Performance Computing on the Cloud via HPC+Cloud software framework. In *Proceedings on 5th International Conference on Eco-Friendly Computing and Communication Systems, ICECCS 2016*, IEEE, , pp. 48–52.

Beloglazov, A., Abawajy, J., Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Generation Computer Systems*, **28**(5), 755–768.

Bonzi, M. (2021). ASPEN GLOBAL CHANGE INSTITUTE ENERGY PROJECT October 2021 Research Review. Retrieved from https://policycommons.net/artifacts/2186496/aspen-global-change-institute-energy-project-october-2021-research-review/2942473/

Buyya, R. (2009). Market-oriented cloud computing: Vision, hype, and reality of delivering computing as the 5th utility. *2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGRID 2009*, **25**(6), 1.

Cidon, A., Stutsman, R., Rumble, S., Katti, S., Ousterhout, J., Rosenblum, M. (2013). MinCopysets: Derandomizing Replication In Cloud Storage. In *Proc. 10th USENIX Symp. Networked Systems Design and Impementation (NSDI)*, , pp. 1–5.

Ding, Y., Qin, X., Liu, L., Wang, T. (2015). Energy efficient scheduling of virtual machines in cloud with deadline constraint. *Future Generation Computer Systems*, **50**, 62–74.

Edwin, E. B., Umamaheswari, P., Thanka, M. R. (2019). An efficient and improved multi-objective optimized replication management with dynamic and cost aware strategies in cloud computing data center. *Cluster Computing*, **22**, 11119–11128.

Fan, C., Jiang, Y., Mostafavi, A. (2021). The Role of Local Influential Users in Spread of Situational Crisis Information. *Journal of Computer-Mediated Communication*, **26**(2), 108–127.

Fan, X., Weber, W. D., Barroso, L. A. (2007). Power provisioning for a warehouse-sized computer. *Proceedings - International Symposium on Computer Architecture*, **35**(2), 13–23.

Hastie, T., Rosset, S., Zhu, J., Zou, H. (2009). Multi-class adaboost. *Statistics and Its Interface*, **2**(3), 349–360.

Huang, Y., Huang, J., Liu, C., Zhang, C. (2020). PFPMine: A parallel approach for discovering interacting data entities in data-intensive cloud workflows. *Future Generation Computer Systems*, **113**, 474–487.

Jahangir, M. H., Mokhtari, R., Mousavi, S. A. (2021). Performance evaluation and financial analysis of applying hybrid renewable systems in cooling unit of data centers – A case study. *Sustainable Energy Technologies and Assessments*, **46**, 101220.

Jennings, B., Stadler, R. (2015). Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management*, **23**(3), 567–619.

Kappelman, L., Torres, R., McLean, E. R., … Guerra, K. (2022). The 2021 SIM IT Issues and Trends Study. *MIS Quarterly Executive*, **21**(1), 75–114.

khalili azimi, S. (2019). A bee colony (beehive) based approach for data replication in cloud environments. In *Lecture Notes in Electrical Engineering*, Vol. 480, Springer, , pp. 1039–1052.

Kireev, V. S., Bochkaryov, P. V., Guseva, A. I., Kuznetsov, I. A., Filippov, S. A. (2019). Monitoring System for the Housing and Utility Services Based on the Digital Technologies IIoT, Big Data, Data Mining, Edge and Cloud Computing. In *Communications in Computer and Information Science*, Vol. 1054, Springer, , pp. 193–205.

Ksentini, A., Taleb, T., Messaoudi, F. (2014). A LISP-Based Implementation of Follow Me Cloud. *IEEE Access*, **2**, 1340–1347.

Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., Jiang, G. (2009). Power and performance management of virtualized computing environments via lookahead control. *Cluster Computing*, **12**(1), 1–15.

Mostafa, N. (2020). A Dynamic Approach for Consistency Service in Cloud and Fog Environment. In *2020 5th International Conference on Fog and Mobile Edge Computing, FMEC 2020*, IEEE, , pp. 28–33.

Mytton, D. (2020). Assessing the suitability of the Greenhouse Gas Protocol for calculation of emissions from public cloud computing workloads. *Journal of Cloud Computing*, **9**(1), 1–11.

Nazir, B., Ishaq, F., Shamshirband, S., Chronopoulos, A. (2018). The Impact of the Implementation Cost of Replication in Data Grid Job Scheduling. *Mathematical and Computational Applications*, **23**(2), 28.

Pierson, J. M., Hlavacs, H. (2015). Introduction to energy efficiency in large-scale distributed systems. *Large-Scale Distributed Systems and Energy Efficiency: A Holistic View*, 1–15.

Ramanan, M., Vivekanandan, P. (2019). Efficient data integrity and data replication in cloud using stochastic diffusion method. *Cluster Computing*, **22**, 14999–15006.

Roser, M. (2022). AI Timelines: What Do Experts in Artificial Intelligence Expect for the Future? *Singularityhub*. Retrieved from https://singularityhub.com/2022/12/18/ai-timelines-what-do-experts-in-artificial-intelligence-expect-for-the-future/

Ruan, K., Carthy, J., Kechadi, T., Baggili, I. (2013). Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results. *Digital Investigation*, **10**(1), 34–43.

Tagne Fute, E., Nyabeye Pangop, D. K., Tonye, E. (2023). A new hybrid localization approach in wireless sensor networks based on particle swarm optimization and tabu search. *Applied Intelligence*, **53**(7), 7546–7561.

Tahir, M., Sardaraz, M., Mehmood, Z., Muhammad, S. (2021). CryptoGA: a cryptosystem based on genetic algorithm for cloud data security. *Cluster Computing*, **24**(2), 739–752.

Zhang, H., Chen, G., Li, X. (2019). Resource management in cloud computing with optimal pricing policies. *Computer Systems Science and Engineering*, **34**(4), 249–254.

Zhang, T. (2020). A QoS-enhanced data replication service in virtualised cloud environments. *International Journal of Networking and Virtual Organisations*, **22**(1), 1–16.

Received April 29, 2024, accepted August 1, 2024

Baltic J. Modern Computing, Vol. 12 (2024), No. 3, pp. 327–348
https://doi.org/10.22364/bjmc.2024.12.3.07

Accuracy vs Complexity: A Small Scale Dynamic Neural Networks Case

Martynas DUMPIS[1], Dalius NAVAKAUSKAS[2]

[1] Vilnius Gediminas Technical University, Plytines g. 25-234, Vilnius LT-10105, Lithuania
[2] Vilnius Gediminas Technical University, Sauletekio al. 11-1103, Vilnius LT-10223, Lithuania

martynas.dumpis@vilniustech.lt, dalius.navakauskas@vilniustech.lt

ORCID 0009-0003-4335-372X, ORCID 0000-0001-8897-7366

Abstract. This research provides a detailed analysis of small-scale dynamic neural network (NN) models for human activity recognition using data from smartphones. We evaluate eight dynamic NN: Finite Impulse Response (FIRNN), Infinite Impulse Response (IIRNN), Gamma Memory (GMNN), Lattice Ladder (LLNN), Time Delay (TDNN), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRUNN), and Long Short-Term Memory (LSTMNN), utilizing a publicly available dataset from Kaggle. The study focuses on comparing these models in terms of higher accuracy, smallest scale, adaptivity to the task (walking vs running classification), and memory utilization. Different NN architectures and synapse configurations are evaluated by their accuracy and computational complexity. The findings reveal which NN architectures offer the best performance while being the least computationally and memory demanding. Among the models, the IIRNN achieved the highest accuracy at 99.86% in the recognition of specified activities. Additionally, the TDNN model demonstrated impressive performance with 99.27% accuracy while requiring fewer computational resources: 2 binary additions, 2 multiplications, and 2 activation functions.

Keywords: Small-Scale Neural Networks, Time-Varying Signals, Smartphone Sensor Data, Human Activity Recognition; Accelerometer

1 Introduction

Recognition of human activities such as walking and running using smartphone sensor data plays a crucial role in advancing health and fitness applications. This capability holds significant promise for healthcare monitoring, athletic training, and lifestyle management. Additionally, integrating Human Activity Recognition (HAR) with IoT technologies paves the way for innovative smart healthcare solutions, demonstrating the convergence of wearable technology and health monitoring (Gomaa and Khamis, 2023). This synergy is beneficial for developing smart cities and personalized health

monitoring systems, thereby expanding the influence of HAR technologies (Serpush et al., 2022). Furthermore, evaluating cognitive-mental abilities such as reaction times, focus, and anticipation through digital solutions can significantly enhance athletic performance and healthcare outcomes, showcasing the importance of integrating cognitive assessments with physical health monitoring (Butkevičiūtė et al., 2023).

This study focuses on a comparative analysis of small-scale dynamic NN models, evaluating their performance, computational requirements, and memory usage in processing accelerometer data. Unlike studies that concentrate on classifying specific activities such as walking and running, this research aims to compare various NN architectures, including TDNN, FIRNN, IIRNN, GMNN, LLNN, RNN, GRUNN, and LSTMNN. Each network is assessed for its efficacy in accurately processing time-sequenced activity data and its adaptability to the task. *The central question* here addressed is what the least complex dynamic NN architectures are and their features that allow one to accurately enough classify walking and running activities using accelerometer data.

The article is structured as follows: the introduction is followed by a review of related work to situate the research within the existing literature. Section on the proposed investigation technique presents the dataset preparation, NN selection, NN models, their complexity, and training procedure. The investigation results encompass investigation coverage and findings on the four main NNs investigation aspects: recognition accuracy, computational complexity, adaptability to the task, and memory utilization. Finally, the implications of the findings are discussed.

2 Related Work

Time-variant data, characterized by its changes over time, is a critical component in numerous research fields. Studies in this area utilize diverse datasets and applications, including HAR, financial forecasting, audio data analysis, medical data, environmental monitoring, and visual data processing.

HAR uses sensor data to recognize human activities, essential for applications in health monitoring and mobile health apps. This field has grown significantly due to the implementation of various models of NNs that enhance the accuracy and efficiency of activity recognition (Kumar et al., 2024). Financial forecasting involves predicting stock prices and other economic indicators using historical financial data (Li et al., 2022). Audio data applications, such as speech recognition, have has played a pivotal role in enabling us to adapt to novel modes of communication: it not only empowers individuals with disabilities to interact, share knowledge and engage in open conversations, but also holds promise for revolutionizing communication between machines using natural languages (Kasparaitis and Antanavičius, 2023; Al-Fraihat et al., 2024).

In the medical domain, time-variant data includes longitudinal patient health records, sensor data from medical devices, and various diagnostic data. These datasets are used to predict and monitor health conditions such as Alzheimer's Disease, diabetes, heart failure, and Parkinson's Disease (Alhudhaif, 2024; Davidashvilly et al., 2024). Environmental data, including air quality and meteorological data, parking data, and Cal-trans

Traffic Performance Measurement System data, is crucial for monitoring and forecasting purposes (Zhang et al., 2024; Weerakody et al., 2021).

Various mathematical models have been developed to handle time-variant data, each with distinct strengths and limitations. Traditional machine learning models such as generative models (maximum a posteriori, Gaussian mixture or hidden Markov) and standard machine learning models (support vector machine, random forest, k-nearest neighbors), linear and autoregressive models (linear regression, autoregressive, autoregressive moving-average, autoregressive conditional heteroskedasticity) have been extensively used. However, these models often fall short in capturing long-term dependencies due to their lack of memory functions, these models are more sensitive to short-term relationships than long-term dependencies, which can not capture some important recurring features (Hamzacebi et al., 2019; Jiang et al., 2020). Generative models, for example, require expert knowledge and preprocessing of text for Automatic Speech Recognition (ASR), making them less flexible than end-to-end ASR models that rely on paired acoustics and language data (Hari et al., 2017).

In recent years, deep learning techniques, particularly recurrent NNs, have gained prominence due to their ability to handle long-term dependencies in sequential data. RNNs, LSTMNN, and GRUNN have demonstrated superior performance in various applications, including healthcare and finance (Kosar and Barshan, 2023). For instance, Deepcare model, which uses Diabetes and Mental Health patient data, achieved a higher F-score (79) compared to traditional models like support vector machine (66.7) and random forests (71.4) (Pham et al., 2016; Weerakody et al., 2021). LSTMNNs, in particular, are known for avoiding long-term dependency issues, making them highly suitable for tasks requiring the recall of information over extended periods (Pham et al., 2016; Hochreiter and Schmidhuber, 1997).

GRUNNs are shown to be less computationally intensive compared to LSTMNNs due to their simpler architecture, resulting in faster training times. However, LSTMNNs generally achieve better predictive performance (Weerakody et al., 2021). Specifically, LSTMNNs demonstrated superior accuracy in capturing long-term dependencies within the data, making them more effective for complex sequence modeling tasks. Furthermore, innovations like Bi-directional GRUNN have improved the accuracy and reliability of HAR systems (Helmi et al., 2023). Additionally, RNNs have been employed in embedded systems for HAR, utilizing data from accelerometers and other sensors to achieve high accuracy, proving their efficiency in real-time applications within resource-constrained environments (Alessandrini et al., 2021). This study adopts streamlined versions of LSTMNN, GRUNN, and RNN architectures to directly compare their performance in HAR.

Primary models such as TDNN, FIRNN, IIRNN, GMNN, and LLNN have also been utilized for handling time-variant data. These models demonstrated good results by incorporating long-term dependencies of input signals but were not directly compared with latest RNNs, GRUNNs, or LSTMNNs. The inclusion of these primary models in this study allows for a comprehensive comparison against latest architectures. TDNNs, known for their implementation simplicity and ability to remember previous signal input values (Paliwal, 1991). FIRNN and IIRNN models excel in signal processing capabilities, while GMNNs are effective in managing long-range dependencies

in sequences (Lawrence et al., 1995). LLNNs, with their unique lattice-ladder structure, offer flexibility and learning potential in dynamic environments, making them suitable for sound and image processing (Back and Tsoi, 1992; Navakauskas et al., 2014; Navakauskienė et al., 2021).

Convolutional Neural Networks (CNNs) are utilized in computer vision tasks. They are also used in HAR with datasets such as KU-HAR, UCI-HAR, and WISDM, showing good results with accuracies of 96.86%, 93.48%, and 93.89%, respectively (Akter et al., 2023). Additionally, models that mix CNNs with LSTMNNs have been effective in identifying a broad range of activities, reaching an accuracy of 90.89% (Khan et al., 2022). However, CNNs typically have pooling layers for down sampling, usually performing average or max pooling to reduce the feature maps spatial resolution (Dentamaro et al., 2024). Such a complexity of NN architecture makes it not suitable for this research due to small size CNNs inability to handle sequential dependencies effectively.

Transformer models (Vaswani et al., 2017) have shown impressive results in natural language processing and image processing, and are beginning to make progress in time series forecasting applications, but are not included in this study due to their complexity and resource-intensive nature. Transformers struggle with feature extraction in time-series data and have high memory requirements, making them unsuitable for small-scale NNs (Weerakody et al., 2021). Another drawback of typical transformers for very long sequences is their memory intensity. Taks needing 1000 s of timestep are particularly difficult due to their quadratic time complexity, which is higher than that of RNNs (Li et al., 2019). Despite advancements in attention mechanisms to address these issues, RNN-based models remain more practical for this study's scope and objectives.

Analysis of related work shows that from the perspective of dynamic NN models TDNN, FIRNN, IIRNN, GMNN, LLNN, LSTMNN, GRUNN, and RNN already are or have potential to be employed in embedded systems for HAR. Thus, it is important to evaluate their performance in recognition of human activities with high accuracy but keeping architectures at a small-scale. This study fills a gap in the literature by including primary dynamic models and comparing them directly with the latest RNN architectures. The development and optimization of these models contribute to a broader understanding of time-variant data applications, including agricultural monitoring and the integration of IoT systems for real-time data processing (Laktionov et al., 2023).

3 Proposed Investigation Technique

This section describes the methodology used for dataset preparation, the selection of dynamic NN models, the specifics and complexity of models, and their training process. The approach is structured to assess the performance of various NN architectures in classifying time-varying human activity data collected by smartphone sensors.

3.1 Dataset Preparation

The primary dataset used in this study is the KU-HAR dataset, obtained from Kaggle (Sikder and Nahid, 2021). This extensive dataset comprises 18 different activities recorded from 90 participants using smartphone sensors, such as accelerometers and

gyroscopes. For the purposes of our research, we focused on accelerometer data corresponding to walking and running activities, which are crucial for HAR in health and fitness applications.

Dataset Characteristics. The KU-HAR dataset contains 1945 raw activity samples and 20,750 subsamples derived from these, captured in both controlled and uncontrolled settings. The activities range from static postures, like standing and sitting, to dynamic movements, such as walking and running. Each sample was carefully recorded to ensure the precision and reliability of the sensor data.

Data Processing. For this study, we used the time domain samples provided in the dataset, which includes 20,750 subsamples, each representing 3 s of non-overlapping accelerometer data. The sampling rate of 100 Hz was standartized across all samples withing all activities.

Magnitude of accelerometer x, y, and z axis coordinates was computed to prepare the data for input into NNs. Acceleration magnitude was calculated as follows:

$$u(n) = \sqrt{x(n)^2 + y(n)^2 + z(n)^2}. \tag{1}$$

Data Partitioning. The subsamples were split into training, validation, and testing sets with a ratio of 60%, 20%, and 20%. This division ensures that the NN models are thoroughly trained, validated for parameter tuning, and finally evaluated on new, unseen data to assess their generalizability and performance.

3.2 Neural Networks Selection

The selection of dynamic NNs for this study was guided by their distinct abilities to manage time-varying signals and intricate data structures. TDNNs were chosen for their simplicity in implementation and their capability to retain previous signal input values (Paliwal, 1991). FIRNN and IIRNN models were selected due to their strong performance in signal processing tasks. GMNNs were included for their proficiency in handling long-range dependencies in sequences, which is especially advantageous for time-varying signals (Lawrence et al., 1995). LLNNs were selected for their unique lattice-ladder structure, offering enhanced flexibility and learning potential in dynamic environments (Back and Tsoi, 1992). This structure has been widely applied in the analysis of sound and image processing (Navakauskas et al., 2014), and its adaptive capabilities have been demonstrated in the study of complex biological datasets in epigenetics (Navakauskienė et al., 2021). RNNs were incorporated due to their fundamental role in learning sequential dependencies within data, which is crucial for modeling cognitive tasks (McClelland and Rumelhart, 1987). GRUNNs, known for their efficiency in sequence modeling as highlighted by Cho et al. (2014), provide a simplified yet robust approach to temporal data processing. LSTMNNs were chosen for their ability to mitigate long-term dependency issues, making them highly suitable for tasks that require remembering information over long durations (Hochreiter and Schmidhuber, 1997).

3.3 Neural Network Models

The forward propagation in each NN model used in this study is specifically tailored to the experimental conditions and architectures employed here, rather than representing generic models. This customization is essential for understanding their operational mechanisms and predictive capabilities. In the following mathematical expressions describing NN models we denote by N_{I} the total number of NN inputs; N_{D} – the order of NN synapse (memory, filter, recurrency, etc.); N_{H} – the total number of NN hidden neurons; Φ_{HT} – the hyperbolic tangent activation function; Φ_{LS} – the logistic sigmoid activation function; $s^{(l)}(n)$ – the l-th layer recall at the time instance n (note, that $s^{(0)}(n)$ is NN input obtained by (1)).

The Output Layer Recall. For all eight selected NN models the output layer recall is calculated similarly and can be expressed by:

$$s^{(2)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{h=1}^{N_{\mathrm{H}}} w_h^{(2)} s_h^{(1)}(n) - \bar{w}^{(2)}\right), \tag{2}$$

here $\boldsymbol{w}^{(2)}$ and $\bar{w}^{(2)}$ – the output neuron layer N_{H} weights vector and a single bias.

The Recall of Hidden Neurons. For each considered NN model the recall of hidden neurons is calculated separately and is provided below.

The hidden neurons of TDNN as synapses use time delay filters, thus the h-th hidden neuron recall is:

$$s_h^{(1)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{i=1}^{N_{\mathrm{I}}} w_{ih}^{(1)} s^{(0)}(n-i) - \bar{w}_h^{(1)}\right), \tag{3}$$

here $\boldsymbol{W}^{(1)}$ and $\bar{\boldsymbol{w}}^{(1)}$ – $N_{\mathrm{I}} \times N_{\mathrm{H}}$ size weights matrix and N_{H} biases vector.

The hidden neurons of FIRNN as synapses use finite impulse response filters $\boldsymbol{w}_{ih}^{(1)}$, thus h-th hidden neuron recall is:

$$s_h^{(1)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{i=1}^{N_{\mathrm{I}}}\sum_{j=1}^{N_{\mathrm{D}}} w_{ijh}^{(1)} s^{(0)}(n-j-i) - \bar{w}_h^{(1)}\right), \tag{4}$$

here $\boldsymbol{W}^{(1)}$ and $\bar{\boldsymbol{w}}^{(1)}$ – $N_{\mathrm{I}} \times N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrix and N_{H} biases vector.

The hidden neurons of IIRNN as synapses use infinite impulse response filters, thus the h-th hidden neuron recall is:

$$s_h^{(1)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{i=1}^{N_{\mathrm{I}}} s_{ih}^{(1)}(n) - \bar{w}_h^{(1)}\right); \tag{5a}$$

$$s_{ih}^{(1)}(n) = \sum_{j=0}^{N_{\mathrm{D}}} b_{ijh}^{(1)}(n) s_i^{(0)}(n-j) + \sum_{j=1}^{N_{\mathrm{D}}} a_{ijh}^{(1)}(n) s_{ih}^{(1)}(n-j), \tag{5b}$$

here $s_{ih}^{(1)}(n)$ is the output of IIR filter; $b_{ijh}^{(1)}(n)$, $a_{ijh}^{(1)}(n)$ are coefficients of feedforward and recursive parts of IIR filter correspondingly; $\boldsymbol{W}^{(1)}$ and $\bar{\boldsymbol{w}}^{(1)}$ – $N_{\mathrm{I}} \times (2N_{\mathrm{D}}+1) \times N_{\mathrm{H}}$ size weights matrix and N_{H} biases vector.

The hidden neurons of GMNN as synapses use Gamma Memory (tuned by $\eta_{ih}^{(1)}$), thus h-th hidden neuron recall is:

$$s_h^{(1)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{i=1}^{N_{\mathrm{I}}}\sum_{j=0}^{N_{\mathrm{D}}} w_{ijh}^{(1)} s_{ijh}^{(0)}(n) - \bar{w}_h^{(1)}\right); \tag{6a}$$

$$s_{ijh}^{(0)}(n) = \begin{cases} s^{(0)}(n-i-1), & j=0; \\ \eta_{ih}^{(1)} s_{i(j-1)h}^{(0)}(n) + \left(1-\eta_{ih}^{(1)}\right) s_{ijh}^{(0)}(n-1), & j \in [1, N_{\mathrm{D}}], \end{cases} \tag{6b}$$

here $s_{ijh}^{(0)}(n)$ – the output signal of the j-th tap of the Gamma Memory connecting i-th input with h-th neuron; $\boldsymbol{W}^{(1)}$ and $\bar{\boldsymbol{w}}^{(1)}$ – $N_{\mathrm{I}} \times (N_{\mathrm{D}}+1) \times N_{\mathrm{H}}$ size weights matrix and N_{H} biases vector.

The hidden neurons of LLNN as synapses use lattice-ladder filters (controlled by lattice $\boldsymbol{k}_{ih}^{(1)}$ and ladder $\boldsymbol{v}_{ih}^{(1)}$ parameters):

$$s_h^{(1)}(n) = \Phi_{\mathrm{LS}}\left(\sum_{i=1}^{N_{\mathrm{I}}}\sum_{j=1}^{N_{\mathrm{D}}} v_{ijh} b_{ijh}^{(1)}(n) - \bar{v}_h^{(1)}\right); \tag{7a}$$

$$\begin{cases} f_{ijh}^{(1)}(n) = f_{i(j-1)h}^{(1)}(n) + k_{ijh}^{(1)} b_{i(j-1)h}^{(1)}(n-1); \\ b_{ijh}^{(1)}(n) = b_{i(j-1)h}^{(1)}(n-1) - k_{ijh}^{(1)} f_{i(j-1)h}^{(1)}(n), \end{cases} \tag{7b}$$

for $j \in [1, N_{\mathrm{D}}]$, with such initial and boundary conditions

$$b_{i0h}^{(1)}(n) = f_{i0h}^{(1)}(n), \qquad f_{i(N_{\mathrm{D}}-1)h}^{(1)}(n) = s^{(0)}(n-i-1). \tag{7c}$$

Here $f_{ijh}^{(1)}(n)$ and $b_{ijh}^{(1)}(n)$ – the lattice forward and errors of backward prediction at the j-th tap, respectively; $\boldsymbol{K}^{(1)}$ – lattice $N_{\mathrm{I}} \times N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrix; $\boldsymbol{V}^{(1)}$ – ladder $N_{\mathrm{I}} \times N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrix; $\bar{\boldsymbol{v}}^{(1)}$ – N_{H} biases vector.

The hidden neurons of RNN together with forward connections use recurrent ones that store neurons hidden states:

$$s_h^{(1)}(n) = \Phi_{\mathrm{HT}}\left(\sum_{i=1}^{N_{\mathrm{I}}} w_{ih}^{(1)} s^{(0)}(n-i-1) + \sum_{j=1}^{N_{\mathrm{D}}} k_{jh}^{(1)} s_h^{(1)}(n-j) - \bar{w}_h^{(1)}\right), \tag{8}$$

here $\boldsymbol{W}^{(1)}$ – forward $N_{\mathrm{I}} \times N_{\mathrm{H}}$ size weights matrix and $\bar{\boldsymbol{w}}^{(1)}$ – N_{H} biases vector; $\boldsymbol{K}^{(1)}$ – recurrent $N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrix.

The neuron's state in *the hidden neurons of GRU* is changed with a candidate state $\tilde{s}_h^{(1)}(n)$, which is updated using the reset gate $s_{\mathrm{R}h}^{(1)}(n)$ and update gate $s_{\mathrm{U}h}^{(1)}(n)$ signals,

thus h-th hidden neuron recall is expressed by:

$$s_h^{(1)}(n) = s_{\mathrm{U}h}^{(1)}(n) s_h^{(1)}(n-1) + \left(1 - s_{\mathrm{U}h}^{(1)}(n)\right) \tilde{s}_h^{(1)}(n); \tag{9a}$$

$$\tilde{s}_h^{(1)}(n) = \Phi_{\mathrm{HT}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) s_{\mathrm{R}h}^{(1)}(n) \right) - \bar{w}_h^{(1)} \right); \tag{9b}$$

$$s_{\mathrm{R}h}^{(1)}(n) = \Phi_{\mathrm{LS}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{R}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{R}h}^{(1)} \right); \tag{9c}$$

$$s_{\mathrm{U}h}^{(1)}(n) = \Phi_{\mathrm{LS}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{U}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{U}h}^{(1)} \right), \tag{9d}$$

here update gate, reset gate and candidate state are represented by: $\boldsymbol{W}_{\mathrm{U}}^{(1)}$, $\boldsymbol{W}_{\mathrm{R}}^{(1)}$, $\boldsymbol{W}^{(1)}$ – combined input and hidden state $N_{\mathrm{I}} \times N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrixes and $\bar{\boldsymbol{w}}_{\mathrm{U}}^{(1)}$, $\bar{\boldsymbol{w}}_{\mathrm{R}}^{(1)}$ and $\bar{\boldsymbol{w}}^{(1)}$ – N_{H} biases vectors, correspondingly.

The hidden neurons of LSTM is support gating of the hidden state. Input gate $s_{\mathrm{I}h}^{(1)}(n)$, forget gate $s_{\mathrm{F}h}^{(1)}(n)$, output gate $s_{\mathrm{O}h}^{(1)}(n)$, and input node $\tilde{s}_{\mathrm{C}h}^{(1)}(n)$ signals are used to construct memory cell internal state $s_{\mathrm{C}h}^{(1)}(n)$. Thus h-th hidden neuron recall is expressed by:

$$s_h^{(1)}(n) = s_{\mathrm{O}h}^{(1)}(n) \Phi_{\mathrm{HT}} \left(s_{\mathrm{C}h}^{(1)}(n) \right); \tag{10a}$$

$$s_{\mathrm{C}ijh}^{(1)}(n) = s_{\mathrm{F}h}^{(1)}(n) s_{\mathrm{C}h}^{(1)}(n-1) + s_{\mathrm{I}h}^{(1)}(n) \tilde{s}_{\mathrm{C}h}^{(1)}(n); \tag{10b}$$

$$\tilde{s}_{\mathrm{C}h}^{(1)}(n) = \Phi_{\mathrm{HT}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{C}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{C}h}^{(1)} \right); \tag{10c}$$

$$s_{\mathrm{I}h}^{(1)}(n) = \Phi_{\mathrm{LS}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{I}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{I}h}^{(1)} \right); \tag{10d}$$

$$s_{\mathrm{F}h}^{(1)}(n) = \Phi_{\mathrm{LS}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{F}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{F}h}^{(1)} \right); \tag{10e}$$

$$s_{\mathrm{O}h}^{(1)}(n) = \Phi_{\mathrm{LS}} \left(\sum_{i=1}^{N_{\mathrm{I}}} \sum_{j=1}^{N_{\mathrm{D}}} w_{\mathrm{O}ijh}^{(1)} \left(s^{(0)}(n-i-1) + s_h^{(1)}(n-j) \right) - \bar{w}_{\mathrm{O}h}^{(1)} \right), \tag{10f}$$

here input, forget, output gate and candidate memory state are denoted by: $\boldsymbol{W}_{\mathrm{I}}^{(1)}$, $\boldsymbol{W}_{\mathrm{F}}^{(1)}$, $\boldsymbol{W}_{\mathrm{O}}^{(1)}$, $\boldsymbol{W}_{\mathrm{C}}^{(1)}$ – combined input and hidden state $N_{\mathrm{I}} \times N_{\mathrm{D}} \times N_{\mathrm{H}}$ size weights matrixes and $\bar{\boldsymbol{w}}_{\mathrm{I}}^{(1)}$, $\bar{\boldsymbol{w}}_{\mathrm{F}}^{(1)}$, $\bar{\boldsymbol{w}}_{\mathrm{O}}^{(1)}$ and $\bar{\boldsymbol{w}}_{\mathrm{C}}^{(1)}$ – N_{H} biases vectors, correspondingly.

3.4 Neural Networks Complexity

Table 1 provides an overview of the computational demands for eight selected NN models, detailing their operational requirements based on the total number of inputs (N_{I}), hidden neurons (N_{H}), and synapse order (N_{D}). The complexity of these dynamic NNs is illustrated by the total number of basic computing elements: bin. additions ($N_{2\Sigma}$), bin. multiplications ($N_{2\Pi}$), and act. functions (N_{Φ}). The highest values in each category are highlighted with a gray background.

Table 1. Dynamic NNs Complexity in Terms of the Total Number of Basic Computing Elements

NN Model	Total Number of Parameters*		
	$N_{2\Sigma}$	$N_{2\Pi}$	N_{Φ}
TDNN	$N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{H}} + 1$
FIRNN	$N_{\mathrm{I}}N_{\mathrm{D}} + N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{I}}N_{\mathrm{D}} + N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{H}} + 1$
IIRNN	$N_{\mathrm{I}}N_{\mathrm{H}}(2N_{\mathrm{D}} + 1) + 2N_{\mathrm{H}}$	$2N_{\mathrm{I}}N_{\mathrm{H}}N_{\mathrm{D}} + N_{\mathrm{H}}$	$N_{\mathrm{H}} + 1$
GMNN	$N_{\mathrm{I}}N_{\mathrm{H}}\left(N_{\mathrm{D}} + \frac{1}{2}(N_{\mathrm{D}}^2 - N_{\mathrm{D}})\right) + N_{\mathrm{I}}(N_{\mathrm{H}} - 1) + 2N_{\mathrm{H}}$	$N_{\mathrm{I}}N_{\mathrm{H}}\left(\frac{1}{2}(N_{\mathrm{D}}^2 + N_{\mathrm{D}}) + 1\right)$	$N_{\mathrm{H}} + 1$
LLNN	$2N_{\mathrm{I}}N_{\mathrm{H}}N_{\mathrm{D}} + 2N_{\mathrm{H}}$	$N_{\mathrm{I}}N_{\mathrm{H}}(2N_{\mathrm{D}} + 2)$	$N_{\mathrm{H}} + 1$
RNN	$N_{\mathrm{I}}N_{\mathrm{D}} + N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{I}}N_{\mathrm{D}} + N_{\mathrm{I}}N_{\mathrm{H}} + N_{\mathrm{H}}$	$N_{\mathrm{H}} + 1$
GRUNN	$3N_{\mathrm{H}}N_{\mathrm{D}}(N_{\mathrm{I}} + N_{\mathrm{D}}) + N_{\mathrm{D}}$	$3N_{\mathrm{I}}N_{\mathrm{H}}(N_{\mathrm{I}} + N_{\mathrm{D}}N_{\mathrm{H}})$	$4N_{\mathrm{D}}N_{\mathrm{H}} + 1$
LSTMNN	$4N_{\mathrm{H}}N_{\mathrm{D}}(N_{\mathrm{I}} + N_{\mathrm{D}}) + N_{\mathrm{D}}$	$4N_{\mathrm{I}}N_{\mathrm{D}}(N_{\mathrm{I}} + N_{\mathrm{D}}N_{\mathrm{H}})$	$6N_{\mathrm{D}}N_{\mathrm{H}} + 1$

* By the gray background, the biggest values of $N_{2\Sigma}$, $N_{2\Pi}$ and N_{Φ} are outlined.

When examining the number of bin. multiplications and bin. additions, GMNN stands out for its complexity, reflecting its design to manage detailed temporal data, which may lead to superior performance in tasks requiring comprehensive historical trend analysis. LLNN also shows a high number of operations, particularly in bin. additions ($2N_{\mathrm{I}}N_{\mathrm{H}}N_{\mathrm{D}} + 2N_{\mathrm{H}}$), highlighting its ability to process data extensively in both forward and backward passes through the layers.

TDNN is characterized by its relatively low complexity, with bin. additions and bin. multiplications scaling linearly with the number of inputs and hidden neurons. This makes TDNN a good choice for simpler, real-time applications. FIRNN, with its higher order synapse filters, introduces more complexity than TDNN but remains efficient in terms of binary operations, making it suitable for tasks requiring finer temporal resolution. IIRNN, while more complex due to its infinite impulse response filters, balances its computational load with enhanced capability to model long-term dependencies, offering a middle ground between simplicity and detailed sequence handling.

The LSTMNN uses the highest number of act. functions ($6N_{\mathrm{D}}N_{\mathrm{H}}+1$), indicating its intricate design aimed at effective memory management. GRUNN, on the other hand, uses fewer act. functions ($4N_{\mathrm{D}}N_{\mathrm{H}}+1$), simplifying some aspects of LSTM's complexity while still maintaining strong sequence processing capabilities. GMNN, LLNN, and RNN utilize the same number of act. functions ($N_{\mathrm{H}} + 1$), suggesting these models

are more straightforward and potentially more suitable for real-time applications like smartphone-based human activity recognition.

3.5 Neural Networks Training

The various NNs in this study were trained using gradient descent methods specifically tailored to their architectures.

TDNN introduces time delays deeper into the structure of NN, requiring modifications to standard NN training algorithms. As proposed by Waibel et al. (1989), the backpropagation (BP) algorithm is adapted for this purpose.

FIRNN training utilizes a variant of the BP algorithm specifically adapted for finite impulse response filters, referred to as temporal BP with gradient descent (Wan, 1990). The training involves updating FIR weights based on the gradient descent method, accounting for the delay elements within the network.

IIRNN training employs BP through time (BPTT) specifically tailored for IIRNN (Campolucci et al., 1999). This method involves unrolling the network through time for each sequence and updating the feedforward weights and IIR filter coefficients based on the gradients calculated throughout this temporal expansion, effectively handling the recursive components of the network.

GMNN employs stochastic gradient descent with BP, with updates applied to both network weights and the Gamma memory parameters (de Vries and Principe, 1992).

LLNN utilizes a simplified stochastic gradient descent with temporal BP, accounting for its lattice-ladder synapse structure (Navakauskas et al., 2014).

RNN uses BPTT (Werbos, 1990), unrolling the network through time for each sequence to update weights based on gradients computed across this temporal expansion.

Both *LSTMNN* and *GRUNN* also use BPTT (Vlachas et al., 2020). LSTMNN incorporates BP through structures that consist input, forget, and output gates, whereas GRUNN uses a similar approach but with simplified update and reset gates.

The Glorot/Xavier *weight initialization* method was adopted to maintain consistent training conditions and optimize weight scaling across different network architectures (Glorot and Bengio, 2010). For initializing hidden layer weights, it is:

$$\boldsymbol{W} = 2r\,\mathrm{rand}(N_\mathrm{I}, N_\mathrm{H}) - r; \tag{11a}$$

$$r = \sqrt{\frac{6}{N_\mathrm{I} + N_\mathrm{H}}}, \tag{11b}$$

here $\mathrm{rand}(\cdot)$ – a matrix of random numbers, in the range $[0,\ 1]$, generator; r – the range for the uniform distribution.

The training process was halted upon meeting *early-stopping criteria* such as when the maximum number of epochs was reached or when the validation loss ceased to improve over several epochs (Prechelt, 1998).

4 Investigation Results

This section presents the investigation coverage and the results of NNs accuracy and complexity evaluation in human activity recognition tasks.

4.1 Investigation Coverage

During the investigation, the size of the NN models was varied while maintaining a fixed three-layer setup.

For FIRNN, IIRNN, LLNN and GMNN, the number of inputs N_{I} ranged from 1 to 10, the synapse order N_{D} ranged from 1 to 10, and the number of hidden neurons N_{H} ranged from 1 to 10, resulting in $10 \times 10 \times 10 = 1000$ architectures for each model.

For TDNN, GRUNN, LSTMNN, and RNN, the synapse (recurrency) order N_{D} varied from 1 to 10, and the number of hidden neurons N_{H} ranged from 1 to 10, resulting in $10 \times 10 = 100$ architectures for each model.

Each NN architecture was initialized and trained 100 times to avoid suboptimal local minima. The architecture with the highest accuracy of these trials was used for further evaluation. In total, this resulted in 440,000 different NN implementations.

To analyze how training duration of various NN models differ, we investigated the distribution of the number of epochs required for each NN model to achieve the highest accuracy as shown in Fig. 1. This analysis was conducted over 100 training runs for each NN model. The IIRNN shows a high median around 140 epochs with significant variability, indicated by a tall box, suggesting less efficient training. The RNN has a low median of 3 epochs, with minimal spread and few outliers, highlighting its efficiency and stable training performance. Similarly, the LSTMNN exhibits a low median of 4 epochs with limited variability, reinforcing its efficient training process. The GRUNN shows the highest median around 180 epochs, indicating it requires the most epochs for training, accompanied by substantial variability and numerous outliers, reflecting poor training efficiency.

The TDNN has a median of 4 epochs with slightly higher variability than RNN and LSTMNN, yet still demonstrates efficient training. The FIRNN displays a low median of 5 epochs with slight variability, maintaining its status as one of the efficient networks. The LLNN shows a higher median around 10 epochs with considerable variability and many outliers, indicating less efficient training compared to others. Lastly, the GMNN

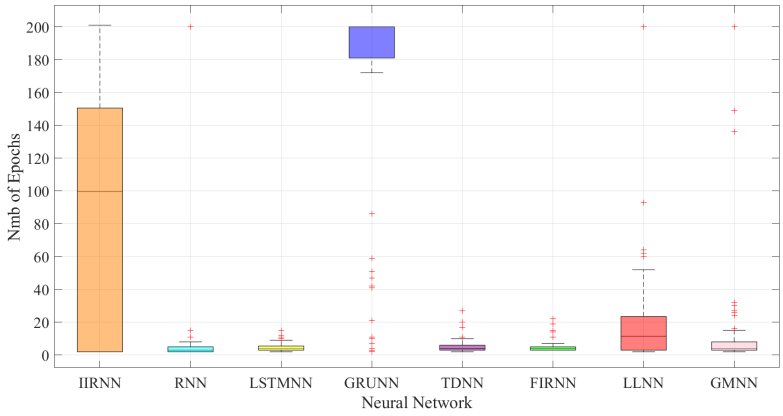


Fig. 1. Training duration analysis of eight dynamic NN models achieving highest accuracy during 100 runs.

network has a low median of 4 epochs with a compressed box towards the bottom, suggesting stable and efficient training performance.

4.2 Neural Networks Evaluation

A comprehensive evaluation of various NN models utilized for HAR was performed. Investigation was focused on the four main aspects: recognition accuracy, computational complexity, adaptability to the task, and memory utilization. Key assessment metrics included accuracy, the number of bin. additions, bin. multiplications, act. functions, weights, and delays, inputs, hidden neurons, and memory allocation size.

Highest Accuracy Neural Networks. Table 2 presents only those NNs that attained the highest accuracy across all architectures within each model. Some models having several entries with the same highest accuracy value are distinguished by additional subscripts. *The main objective* of the analysis is to identify architectures that not only achieve top accuracy (rows in the table are primarily sorted in decreasing accuracy order) but also balance computational efficiency, minimizing the number of bin. additions, bin. multiplications, and act. functions (complementary sorting of rows in the table in increasing number of parameters order).

Table 2. Highest Accuracy Achieving Dynamic Neural Networks and Their Complexity

Neural Network*	Total Number of Parameters**					Highest ACC, %
	$N_{2\Sigma}$	$N_{2\Pi}$	N_{Φ}	N_{W}	N_{D}	
IIRNN_1	612	549	10	5	603	99.86
IIRNN_2	644	595	8	5	637	99.86
IIRNN_3	920	850	11	7	910	99.86
RNN_1	14	14	5	6	44	99.77
RNN_2	19	19	7	7	84	99.77
TDNN_1	35	35	6	6	35	99.77
TDNN_2	40	40	6	7	40	99.77
TDNN_3	49	49	8	6	49	99.77
TDNN_4	63	63	10	6	63	99.77
GMNN	1503	1512	4	10	597	99.75
LLNN_1	1638	1782	10	10	189	99.75
LLNN_2	1820	1980	11	10	210	99.75
LSTMNN	490	480	61	10	490	99.66
FIRNN	35	35	6	5	80	99.58
GRUNN_1	196	189	29	7	196	99.55
GRUNN_2	306	297	37	9	306	99.55

* The gray background outlines the NNs with highest ACC or lowest parameter values.
** $N_{2\Sigma}$ – bin. additions; $N_{2\Pi}$ – bin. multiplications; N_{Φ} – act. func.; N_{W} – weights; N_{D} – delays.

IIRNN_1, IIRNN_2, and IIRNN_3 all achieved the highest accuracy of 99.86%. However, they required a significant amount of computational resources. The least resource-

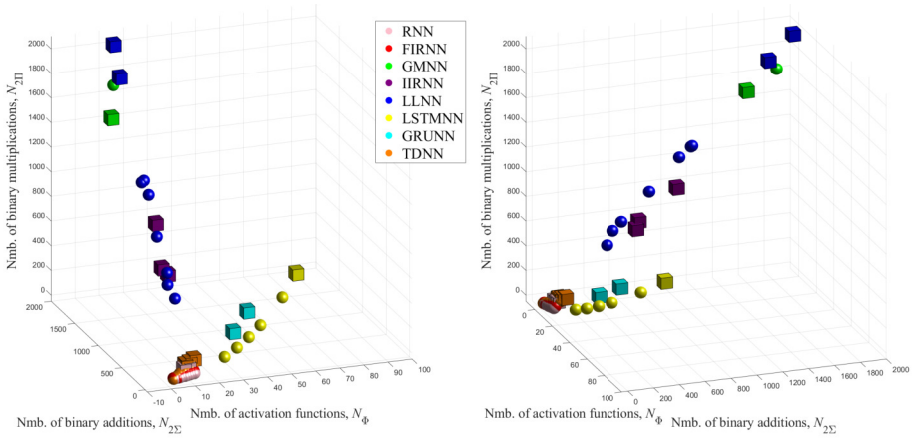
intensive of the three still required 604 bin. multiplications, 549 bin. additions, and 10 act. functions. Although RNN_1 required the fewest bin. multiplications (14) and bin. additions (14) and achieved the second highest accuracy (99.77%). Although RNN_2 had similar accuracy, it needed marginally more computational resources, with 19 bin. multiplications and 19 bin. additions. GMNN and both LLNN configurations also reached high accuracy at 99.75%. Despite GMNN having the fewest act. functions (4), it required a large number of other parameters, with 1512 bin. multiplications and 1503 bin. additions, indicating significant computational demand. $LLNN_1$ and $LLNN_2$ also demanded extensive computational resources, with 1782 and 1980 bin. multiplications, and 1638 and 1820 bin. additions, respectively, demonstrating a trade-off between accuracy and computational load. Notably, LSTMNN achieved a high performance with an accuracy of 99.66%, but it required the most act. functions, totaling 61.

Close to the Highest Accuracy Neural Networks. Computational complexity data of an expanded range of NN architectures is presented in Fig. 2 on the next page. *The main objective* of the analysis is to relax the demands on accuracy (lowering acceptable accuracy level or looking for the simplest complexity NN architecture) to get insights on trade-off between computational complexity and accuracy across investigated NN models.

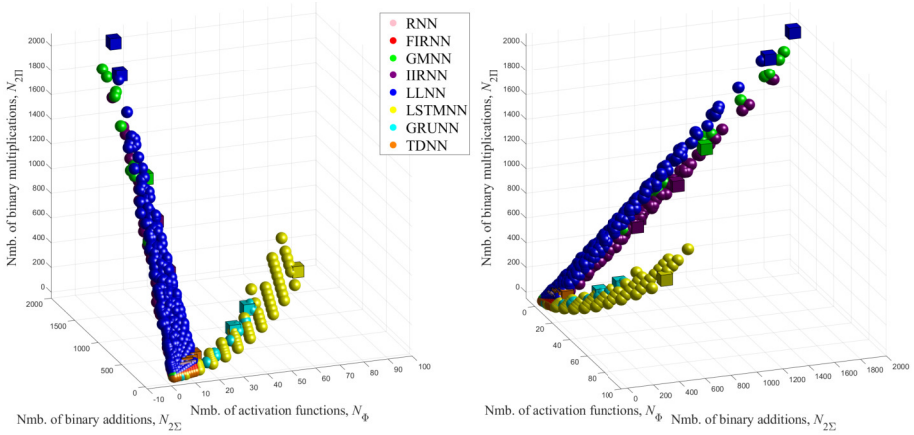
The graph on the lest side in Fig. 2(a) primarily focuses on the computational complexity between the number of act. functions and bin. multiplication operations across various NN models, while also showing bin. additions in a 3D perspective. It includes 19 TDNN, 24 FIRNN, 3 IIRNN, 2 GMNN, 9 LLNN, 22 RNN, 6 LSTMNN, and 2 GRUNN architectures. Cubes denote the NN architectures with the highest accuracy, while spheres represent those within 99.9% of the highest accuracy attained by the leading $IIRNN_1$ network. A dense cluster of FIRNNs, TDNNs and RNNs at the bottom of the graph indicates these networks require fewer act. functions. IIRNN, GMNN, and LLNN exhibit a moderate number of act. functions, with differing bin. multiplication needs. GRUNN shows a moderate increase in both bin. multiplications and act. functions, indicating a higher but manageable computational load. However, LSTMNNs are characterized by the highest number of act. functions.

The graph on the right side in Fig. 2(a) focuses on the computational complexity between the number of bin. multiplications and bin. additions, while also showing act. functions in a 3D perspective. LLNN shows a noticeable linear relationship, with an increase in bin. multiplications generally corresponding to an increase in bin. additions. FIRNN, TDNN, and RNN configurations cluster towards the lower end of the graph, while GRUNN and LSTMNN occupy the middle range, with GRUNN showing slightly lower computational needs than LSTMNN. In contrast, IIRNN, and LLNN are spread across the range, indicating variable computational requirements. GMNN stands out, reflecting the highest demands for both bin. multiplications and additions.

To explore a broader spectrum of our implemented NN architectures, we established a new threshold of 99% accuracy across all NN models. This threshold enables us to investigate deeper relationships between computational parameters and identify smaller, less resource-intensive architectures that maintain high accuracy.



(a) Neural network architectures achieving an accuracy rate of 99.9% of the highest accuracy attained by the leading IIRNN_1 network



(b) Neural network architectures achieving an accuracy rate of 99% or higher

Fig. 2. Computational complexity analysis of an expanded range of NN architectures. The perspectives offered include: on the left side – the relationship between bin. multiplications and act. functions, and on the right side – the relationship between bin. multiplications and bin. additions. Cubes (instead of spheres) denote the NN architectures with the highest accuracy.

The Fig. 2(b) expands the analysis to a wider range of NN models (all achieving an accuracy rate of 99% or higher): 71 TDNN, 248 FIRNN, 142 IIRNN, 55 GMNN, 360 LLNN, 92 RNN, 58 LSTMNN, and 15 GRUNN. This broader dataset reveals that the linear dependency between the number of bin. multiplications and bin. additions is not exclusive to LLNN; it is also evident in FIRNN, GMNN, IIRNN, LSTMNN, and GRUNN. Furthermore, a linear trend is observed for LSTMNN and GRUNN in terms of act. functions with respect to both bin. additions and bin. multiplications. This suggests a more generalizable relationship across different NN architectures, highlight-

ing consistent computational patterns and dependencies that could inform optimization strategies for various NN designs.

Finally, to visualize not only the best accuracy achieving NNs but also the smallest architecture NNs, with the current threshold set at 99% accuracy we looked for the smallest architectures (sorting then according to the smallest number of bin. additions, bin. multiplications, act. functions, weights, and delays). Table 3 provides a comprehensive comparison of NN structures that achieve an accuracy of 99% or higher while maintaining minimal computational complexity within each NN model.

Table 3. Neural Network Architectures with Minimal Computational Complexity Achieving 99% or Higher Accuracy

Neural Network*	Total Number of Parameters**					Highest ACC, %
	$N_{2\Sigma}$	$N_{2\Pi}$	N_{Φ}	N_{W}	N_{D}	
TDNN	2	2	2	2	1	99.27
GMNN	3	2	2	5	1	99.35
FIRNN	3	3	2	2	1	99.26
RNN	1	3	2	3	1	99.18
LLNN	6	8	2	3	1	99.27
GRUNN	7	6	5	7	1	99.23
LSTMNN	9	8	7	9	1	99.35
IIRNN	44	34	3	2	42	99.29

* The gray background outlines the NNs with highest ACC or lowest parameter values.
** $N_{2\Sigma}$ – bin. additions; $N_{2\Pi}$ – bin. multiplications; N_{Φ} – act. func.; N_{W} – weights; N_{D} – delays.

The FIRNN model demonstrates exceptional efficiency, achieving an accuracy of 99.27% with the lowest parameter counts: 1 bin. addition, 2 bin. multiplications, 2 act. functions, 2 weights, and 1 delay. The RNN achieves a similar accuracy of 99.18% with slightly higher parameter requirements. TDNN also shows a high accuracy of 99.24% with modest computational complexity. The LLNN model, while achieving 99.27% accuracy, requires more parameters, particularly in bin. additions and multiplications. GRUNN achieves 99.23% accuracy but with a considerable increase in the number of parameters. LSTMNN, achieving the highest accuracy of 99.35%, also requires the highest number of parameters among the simpler networks. Finally, the IIRNN structure, with an accuracy of 99.29%, demands a significant number of parameters, indicating a higher computational complexity. This table highlights the trade-off between computational complexity and accuracy across different NN models.

Adaptability of Neural Networks to the Task. Table 4 presents NNs that achieved 99% or higher accuracy, architectural parameters, specifically the number of delays, inputs, and hidden neurons. *The main objective* of the analysis is to get insights on NN models intrinsic flexibility to adapt to the given HAR task.

For RNNs, it is noticeable that most of the most accurate architectures vary in terms of the number of hidden neurons, ranging from 2 to 10. However, the number of delays

Table 4. Summary of Neural Network Architectures that Achieved 99% or Higher Accuracy

Neural Network	Total Number of Parameters*		
	Inputs, N_I	Delays, N_D	Hidden Neurons, N_H
TDNN	1, 2, 5, 6, 7, 8	1	1 – 5 – 10
FIRNN	1, 2	1 – 5 – 10	1 – 5 – 10
GMNN	1 – 9	1 – 3 – 10	1 – 10
IIRNN	3, 4, 6, 7	2 – 9, 10	4, 5 – 10
LLNN	2 – 9 – 10	1 – 9 – 10	1 – 10
LSTMNN	1 – 2 – 6	1 – 10	1
RNN	1	1 – 4 – 10	2 – 6 – 10
GRUNN	1, 2	1 – 7 – 10	1

* The gray background outlines the architecture parameters of the NNs that achieved the highest accuracy.

varies across the entire tested range (from 1 to 10). The most accurate architecture for RNNs has 1 input, 4 delays, and 6 hidden neurons. In FIRNNs, the optimal choice is to use 1 or 2 hidden neurons, as adding more tends to decrease the network's performance for this task. The number of inputs and delays for FIRNNs does not significantly affect performance, as they are spread across all tested variations (from 1 to 10). The most accurate architecture for FIRNNs has 2 inputs, 5 delays, and 5 hidden neurons.

For GMNNs, it is suggested to use between 1 and 6 hidden neurons. A clear linear relationship between the number of inputs and delays is noticeable, suggesting that using the same number of delays and inputs for a single structure, varying from 1 to 10, is beneficial. The most accurate architecture for GMNNs has 9 inputs, 3 delays, and 10 hidden neurons. For IIRNNs, it is recommended to use 3, 4, 6, or 7 inputs with 4 or more hidden neurons and 2 or more delays. The most accurate architecture for IIRNNs has 6 inputs, 9 delays, and 5 hidden neurons.

LLNNs show that there should be at least 2 inputs, and the number of delays should be equal to or greater than the number of inputs. The number of hidden neurons for LLNNs does not matter much, as the most accurate architectures are equally spread across all tested variations (from 1 to 10). The most accurate architecture for LLNNs has 9 inputs, 9 delays, and 10 hidden neurons. For LSTMNNs, the number of delays is less important than the number of inputs. It is noticeable that when the number of inputs exceeds 6, the number of good architectures decreases significantly. The most accurate architecture for LSTMNNs has 2 inputs, 10 delays, and 1 hidden neuron.

For GRUNNs, using 1 input with 1 to 10 delays is suggested. The most accurate architecture for GRUNNs has 2 inputs, 7 delays, and 1 hidden neuron. For TDNNs, most of the most accurate architectures appear when there are 1, 2, 5, 6, 7, or 8 inputs.

Memory Utilization of Neural Networks. The memory utilization patterns of different NNs architectures are shown in Fig. 3. Each graph illustrates the relationship between memory usage and accuracy (ACC, %) for a specific NN architecture. Each dot in the graphs represents a different configuration of the respective architecture, varying

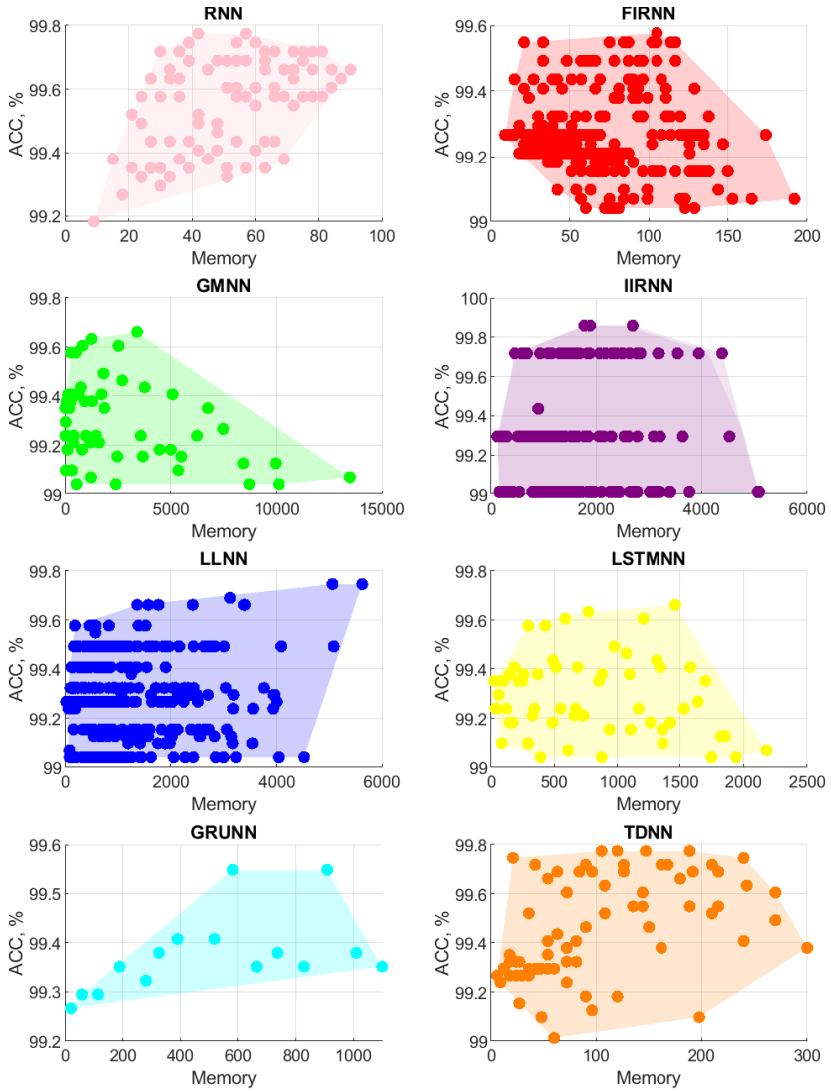


Fig. 3. Memory utilization of dynamic NN models, each achieving at least 99% accuracy. The graph shows the relationship between memory usage (sum of weights, delays, and coefficient-adjusted activation functions) and accuracy across different architectures.

in the number of inputs, delay elements, or hidden neurons. *The main objective* of the analysis is to identify memory allocation needs for each NNs architecture as also as to get insights on memory usage influence on the overall NNs accuracy.

Memory utilization in these NNs consists of several components. The total memory is the sum of the number of weights and the number of delays, with equal memory allocation needed for each weight and delay. Additionally, act. functions contribute to

memory usage. They are evaluated as lookup tables, and their memory contribution is not directly equivalent to weights or delays. Therefore, the memory allocated for act. functions is multiplied by a specific coefficient to account for their different impact on overall memory usage.

In Fig. 3 RNN graph when memory allocation increases from approximately 0 to 80 units, the accuracy generally trends upwards from 99.2 to about 99.8%. This suggests that RNNs benefit from increased memory, enhancing their ability to learn and retain temporal dependencies, thereby improving performance.

The FIRNN graph shows that accuracy peaks around 100 units of memory usage and then declines. Accuracy ranges from 99.0% to 99.6%, indicating an optimal memory size for FIRNNs. Beyond this point, more memory doesn't lead to better accuracy and may even reduce performance. This suggests a critical balance in memory allocation for FIRNNs.

The GMNN graph shows a clear positive correlation between memory and accuracy up to a certain threshold. Memory usage ranges widely from 0 to 15,000 units, with accuracy improving from 99.0% to about 99.8%. This indicates that GMNNs use large memory capacities to manage detailed temporal data effectively, enhancing performance in tasks requiring comprehensive historical analysis.

The IIRNN graph shows a high but stable accuracy range from 99.0% to 99.86%, with memory usage from 0 to 6,000 units. Accuracy doesn't vary much with memory changes, suggesting IIRNNs maintain high performance across different memory levels. This stability indicates efficient memory usage in IIRNNs. A noticeable feature in the IIRNN graph is the distinct horizontal lines formed by data points at specific accuracy levels: 99.0%, 99.25%, and 99.75%. These lines suggest that certain configurations of IIRNNs consistently achieve these accuracy levels regardless of variations in memory usage. This pattern indicates that while memory allocation is crucial, there are other factors within the IIRNN architecture that strongly influence its accuracy, leading to these stable performance bands.

The LLNN graph shows a slight positive trend in accuracy with increasing memory usage, ranging from 0 to 6,000 units. Accuracy improves from 99.0% to about 99.8%, implying LLNNs can benefit from more memory, but less dramatically than other architectures. Moreover, there are noticeable horizontal lines at accuracy levels such as 99.1%, 99.3%, 99.4%, and 99.5%. These lines indicate that LLNNs also have configurations that consistently achieve these accuracy levels. The presence of these lines suggests that while memory usage impacts performance, certain LLNN configurations can maintain specific accuracy thresholds, highlighting a degree of robustness in their design.

The LSTMN graph shows memory usage from 0 to 2,500 units, with accuracy improving from 99.0% to about 99.8%. This positive correlation indicates that LSTMNs effectively use more memory to maintain long-term dependencies, which is crucial for tasks requiring extended temporal context.

The GRUNN graph shows a notable positive correlation between memory usage and accuracy, with memory ranging from 0 to 1,000 units. Accuracy increases from 99.2% to about 99.6%, showing that GRUNNs use more memory to improve performance.

This architecture's ability to manage and update memory states dynamically adds to its efficiency.

The TDNN graph shows a broader distribution of memory usage from 0 to 300 units, with accuracy ranging from 99.2% to 99.8%. There is a slight positive trend, indicating TDNNs benefit from more memory to some extent. However, the variability suggests other factors also affect their accuracy.

5 Conclusions

This study explored eight different models of small-scale dynamic neural networks to identify the least complex architectures capable of accurately classifying walking and running activities using accelerometer data. Based on the analysis of 440,000 distinct NN implementations, the following key points were observed:

1. Among the configurations tested, the IIRNN, especially IIRNN_1, achieved the highest accuracy of 99.86%. with the least computational demand in terms of bin. additions and bin. multiplications.
2. The TDNN demonstrated an excellent balance by providing a high accuracy of 99.27% while requiring lowest computational complexity in terms of bin. additions and bin. multiplications making it highly suitable for real-time applications.
3. Despite its higher computational demands, the GMNN exhibited strong performance, achieving 99.35% accuracy with the fewest number of act. functions (2).
4. Among the NN models analyzed, FIRNNs and TDNNs stand out for their ability to achieve high accuracy with minimal architectural complexity. FIRNNs perform optimally with 1 or 2 hidden neurons, 2 inputs, and 5 delays, while TDNNs excel with 1 or 2 inputs and 6 hidden neurons, making both networks highly suitable for tasks requiring efficient, real-time processing.
5. Notably, the IIRNN and LLNN graphs show distinct horizontal lines at certain accuracy levels, indicating that these networks achieve stable and consistent performance across a range of memory usages. This robustness makes them reliable choices for applications where maintaining high accuracy regardless of memory constraints is essential.

References

Akter, M., Ansary, S., Khan, M. A.-M., Kim, D. (2023). Human activity recognition using attention-mechanism-based deep learning feature combination, *Sensors* **23**(12), 15. https://doi.org/10.3390/s23125715

Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M., Maaita, A. (2024). Speech recognition utilizing deep learning: A systematic review of the latest developments, *Human-Centric Computing and Information Sciences* **14**, 19143 – 19165. https://doi.org/10.22967/HCIS.2024.14.015

Alessandrini, M., Biagetti, G., Crippa, P., Falaschetti, L., Turchetti, C. (2021). Recurrent neural network for human activity recognition in embedded systems using PPG and accelerometer data, *Electronics* **10**(14), 1–18. https://doi.org/10.3390/electronics10141715

Alhudhaif, A. (2024). A novel approach to recognition of alzheimer's and parkinson's diseases: Random subspace ensemble classifier based on deep hybrid features with a super-resolution image, *PeerJ Computer Science* **10**, e1862. https://doi.org/10.7717/peerj-cs.1862

Back, A. D., Tsoi, A. C. (1992). An adaptive lattice architecture for dynamic multilayer perceptrons, *Neural Computation* **4**(6), 922–931. https://doi.org/10.1162/neco.1992.4.6.922

Butkevičiūtė, E., Bikulčienė, L., Blazauskas, T., Žemaitytė, A. (2023). Cognitive checkup and mental training platform for elite athletes, *Baltic Journal of Modern Computing* **11**(2), 257–271. https://doi.org/10.22364/bjmc.2023.11.2.03

Campolucci, P., Uncini, A., Piazza, F., Rao, B. D. (1999). On-line learning algorithms for locally recurrent neural networks, *IEEE Transactions on Neural Networks* **10**(2), 253–271. https://doi.org/10.1109/72.750549

Cho, K., Merrienboer, B., Bahdanau, D., Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches, *Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)* p. 9. https://doi.org/10.3115/v1/W14-4012

Davidashvilly, S., Cardei, M., Hssayeni, M., Chi, C., Ghoraani, B. (2024). Deep neural networks for wearable sensor-based activity recognition in parkinson's disease: Investigating generalizability and model complexity, *BioMedical Engineering OnLine* **23**(1), 17. https://doi.org/10.1186/s12938-024-01214-2

de Vries, B., Principe, J. C. (1992). The gamma model–a new neural model for temporal processing, *Neural Networks* **5**(4), 565–576. https://doi.org/10.1016/S0893-6080(05)80035-8

Dentamaro, V., Gattulli, V., Impedovo, D., Manca, F. (2024). Human activity recognition with smartphone-integrated sensors: A survey, *Expert Systems with Applications* **246**, 1–22. https://doi.org/10.1016/j.eswa.2024.123143

Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks, *in* Teh, Y. W., Titterington, M. (eds), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Vol. 9 of *Proceedings of Machine Learning Research*, PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 249–256.

Gomaa, W., Khamis, M. A. (2023). A perspective on human activity recognition from inertial motion data, *Neural Computing & Applications* p. 20463–20568. https://doi.org/10.1007/s00521-023-08863-9

Hamzacebi, C., Avni, H. E., Cakmak, R. (2019). Forecasting of Turkey's monthly electricity demand by seasonal artificial neural network, *Neural Computing & Applications* **31**(7), 2217–2231. https://doi.org/10.1007/s00521-017-3183-5

Hari, T., Watanabe, S., Zhang, Y., Chan, W. (2017). Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM, *18th Annual Conference of the International Speech Communication Association (Interspeech 2017), vols 1–6: Situated Interaction*, Interspeech, pp. 949–953. https://doi.org/10.21437/Interspeech.2017-1296

Helmi, A. M., Al-qaness, M. A. A., Dahou, A., Abd Elaziz, M. (2023). Human activity recognition using marine predators algorithm with deep learning, *Future Generation Computer Systems-the International Journal of Escience* **142**, 340–350. https://doi.org/10.1016/j.future.2023.01.006

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jiang, P., Li, R., Liu, N., Gao, Y. (2020). A novel composite electricity demand forecasting framework by data processing and optimized support vector machine, *Applied Energy* **260**, 1–15. https://doi.org/10.1016/j.apenergy.2019.114243

Kasparaitis, P., Antanavičius, D. (2023). Investigation of input alphabets of end-to-end lithuanian text-to-speech synthesizer, *Baltic Journal of Modern Computing* **11**(2), 285–296. https://doi.org/10.22364/bjmc.2023.11.2.05

Khan, I. U., Afzal, S., Lee, J. W. (2022). Human activity recognition via hybrid deep learning based model, *Sensors* **22**(1), 1–16. https://doi.org/10.3390/s22010323

Kosar, E., Barshan, B. (2023). A new CNN-LSTM architecture for activity recognition employing wearable motion sensor data: Enabling diverse feature extraction, *Engineering Applications of Artificial Intelligence* **124**, 1–15. https://doi.org/10.1016/j.engappai.2023.106529

Kumar, P., Chauhan, S., Awasthi, L. K. (2024). Human activity recognition (HAR) using deep learning: Review, methodologies, progress and future research directions, *Archives of Computational Methods in Engineering* **31**(1), 179–219. https://doi.org/10.1007/s11831-023-09986-x

Laktionov, I., Diachenko, G., Koval, V., Yevstratiev, M. (2023). Computer-oriented model for network aggregation of measurement data in iot monitoring of soil and climatic parameters of agricultural crop production enterprises, *Baltic Journal of Modern Computing* **11**(3), 500–522. https://doi.org/10.22364/bjmc.2023.11.3.09

Lawrence, S., Tsoi, A., Back, A. (1995). The gamma MLP for speech phoneme recognition, *in* Touretzky, D., Mozer, M., Hasselmo, M. (eds), *Advances in Neural Information Processing Systems*, Vol. 8, MIT Press, pp. 785–791.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, *in* Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R. (eds), *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, Vol. 32 of *Advances in Neural Information Processing Systems*, pp. 1–14. 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, DEC 08-14, 2019.

Li, Z., Liao, Y., Hu, B., Ni, L., Lu, Y. (2022). A financial deep learning framework: Predicting the values of financial time series with ARIMA and LSTM, *International Journal of Web Services Research* **19**(1), 1–15. https://doi.org/10.4018/IJWSR.302640

McClelland, J. L., Rumelhart, D. E. (1987). *Schemata and Sequential Thought Processes in PDP Models*, MIT press. https://doi.org/10.7551/mitpress/5236.003.0004

Navakauskas, D., Serackis, A., Matuzevičius, D., Laptik, R. (2014). *Specializuotos elektroninės intelektualiosios sistemos garsams ir vaizdams apdoroti. Teorija ir taikymai*, Vilnius Gediminas Technical University. https://doi.org/10.3846/2310-m

Navakauskienė, R., Navakauskas, D., Borutinskaitė, V., Matuzevičius, D. (2021). *Epigenetics and Proteomics of Leukemia: A Synergy of Experimental Biology and Computational Informatics*, Springer International Publishing. https://doi.org/10.1007/978-3-030-68708-3

Paliwal, K. (1991). A time-derivative neural net architecture - an alternative to the time-delay neural net architecture, *Neural Networks for Signal Processing Proceedings of the 1991 IEEE Workshop*, pp. 367–375. https://doi.org/10.1109/NNSP.1991.239505

Pham, T., Tran, T., Phung, D., Venkatesh, S. (2016). DeepCare: A deep dynamic memory model for predictive medicine, *Advances in Knowledge Discovery and Data Mining, Pakdd 2016, PT II*, Vol. 9652 of *Lecture Notes in Artificial Intelligence*, pp. 30–41. https://doi.org/10.1007/978-3-319-31750-2_3

Prechelt, L. (1998). Automatic early stopping using cross validation: Quantifying the criteria, *Neural Networks* **11**(4), 761–767. https://doi.org/10.1016/S0893-6080(98)00010-0

Serpush, F., Menhaj, M. B., Masoumi, B., Karasfi, B. (2022). Wearable sensor-based human activity recognition in the smart healthcare system, *Computational Intelligence and Neuroscience* **2022**, 1–23. https://doi.org/10.1155/2022/1391906

Sikder, N., Nahid, A.-A. (2021). KU-HAR: An open dataset for heterogeneous human activity recognition, *Pattern Recognition Letters* **146**, 46–54. https://doi.org/10.1016/j.patrec.2021.02.024

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need, *in* Guyon, I., Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds), *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Vol. 30 of *Advances in Neural Information Processing Systems*,

pp. 1–11. 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, DEC 04-09, 2017.

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., Koumoutsakos, P. (2020). Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics, *Neural Networks* **126**, 191–217. https://doi.org/10.1016/j.neunet.2020.02.016

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K. (1989). Phoneme recognition using time-delay neural networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**(3), 328–339. https://doi.org/10.1109/29.21701

Wan, E. A. (1990). Temporal backpropagation for FIR neural networks, *IJCNN International Joint Conference on Neural Networks, Vols 1–3*, Int Neural Network Soc, pp. A575–A580. International Joint Conf on Neural Networks (IJCNM-90), San Diego, CA, JUN 17-20, 1990.

Weerakody, P. B., Wong, K. W., Wang, G., Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks, *Neurocomputing* **441**, 161–178. https://doi.org/10.1016/j.neucom.2021.02.046

Werbos, P. J. (1990). Backpropagation through time – what it does and how to do it, *Proceedings of the IEEE* **78**(10), 1550–1560. https://doi.org/10.1109/5.58337

Zhang, Z., Zhang, S., Chen, C., Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning, *Alexandria Engineering Journal* **93**, 128–141. https://doi.org/10.1016/j.aej.2024.03.031

Received July 7, 2024 , accepted September 20, 2024