JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)

VOLUME: 3 ISSUE NO: 1 YEAR: 2023 PERIOD: JANUARY-DECEMBER
e-ISSN: 2757-8267

INTERNATIONAL, OPEN ACCESS, PEER-REVIEWED JOURNAL

http://dergipark.org.tr/ject

İZMİR AKADEMİ DERNEĞİ

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)

E-ISSN: 2757-8267

Volume :3 Number:1 Year:2023

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)

is indexed and abstracted by

Index Copernicus

https://journals.indexcopernicus.com/search/details?id=68795

ROAD

https://portal.issn.org/resource/ISSN/2757-8267

Academia.edu

https://gkgkgkg.academia.edu/JournalofEmergingComputerTechnologies

Google Scholar

https://scholar.google.com/citations?user=n986qsIAAAAJ&hl=tr

Asos Index

https://asosindex.com.tr/index.jsp?modul=journal-page&journal-id=2594

Academic Resource Index (Researchbib)

http://journalseeker.researchbib.com/view/issn/2757-8267

OpenAIRE

https://explore.openaire.eu/search/find?fv0=%222757-8267%22&f0=q&active=result

IAD

http://openaccess.izmirakademi.org/index.php/iad/preprints/category/ject

Cosmos

http://www.cosmosimpactfactor.com/page/journals_details/6532.html

EuroPub

https://europub.co.uk/journals/29522

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES

Volume: 3 **No:** 1 **Year:** 2023 **Period:** January-December

Head Office : İzmir
Type of Publication : It is published electronically once a year, between January and December.
Owner : İzmir Akademi Derneği (İzmir Academy Association)

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES

Volume: 3 **No:** 1 **Year:** 2023 **Period:** January-December

Journal of Emerging Computer Technologies

Address : Büyükkale Mh. Kerimoğlu Sok. No:40-2 Tire / Izmir, TURKEY
Web : https://dergipark.org.tr/en/pub/ject
Email : izmirakademidernegi@outlook.com
Email : izmirakademidernegi@gmail.com

Journal of Emerging Computer Technologies publishes scientific/original research articles. It is published electronically twice a year, in June and December. It is an **international scientific refereed journal** that publishes articles written in **English**, and includes acedemicians from different countries in its boards. Uses "Double-Blind Peer Review" in reviewing processes. It has adopted the open access principle. No fee is requested from the authors for open access, processing, publication fee or otherwise. It is totally **FREE.**

CONTENT

Journal of Emerging Computer Technologies
Research Article
Received: *2023-05-17* | ***Reviewing:*** *2023-05-17* & *2023-08-02*| ***Accepted:*** *2023-08-02* | ***Online:*** *2023-08-07* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1297961

What if GPT4 Became Autonomous: The Auto-GPT Project and Use Cases

Mehmet Fırat
Department of Distance Education
Open Education Faculty
Anadolu University
Eskişehir, Türkiye
mfirat@anadolu.edu.tr
0000-0001-8707-5918

Saniye Kuleli
İzmir Provincial Directorate of National Education
İzmir,Türkiye
saniye.kuleli@meb.gov.tr
0000-0001-7838-4997
(*Corresponding Author*)

***Abstract*— Auto-GPT is a product of an experimental project that makes the use of GPT-4 autonomous. Notably, Auto-GPT emerged and spread rapidly, while the echo of OpenAI's ChatGPT continues. However, there are insufficient studies on this new application in related literature. The purpose of this exploratory case study was to explore the different use cases and experiences of Auto-GPT users. For this purpose, 16 users with an Auto-GPT experience on the GitHub platform were interviewed. Thematic content analysis was performed on the qualitative data. AutoGPT experiences of users can be characterized by learning programs, autonomous applications, conducting research, and writing reports. The results of this study showed that content creation is the most important purpose of using Auto-GPT. As independent research functions of Auto-GPT, users also emphasize data summarization and information organization. However, the participants also pointed out the token limit (inefficiency), forgetting generated tools, and iteration as some prominent limitations of Auto-GPT. It is possible to say that Auto-GPT has a high potential to use in also in educational purpose, but it is still in the development stage.**

Keywords— Auto-GPT, GPT-4, ChatGPT, Use Cases

I. INTRODUCTION

What artificial intelligence does and can do has undoubtedly been one of the most important topics of discussion nowadays among technology developers and users. Natural language processing models have become increasingly popular in recent years and are recognized worldwide as an important milestone in the development and use of artificial intelligence [1]. A succession of fast and efficient models reshapes the way we interact with AI and the digital world. In particular, OpenAI's GPT series, with its innovative features, has led to exciting developments in the emergence of opportunities arising from the collaboration of advanced language models with human and artificial intelligence. Natural language processing is a subset of artificial intelligence which refers to the automatic computational processing of human speech, and includes systems which use human-generated text as input and systems which produce natural-sounding text as output [2].

Auto-GPT is a product of an experimental project developed to make the use of GPT-4 autonomous. Auto-GPT is a Python application that is available as an open source on GitHub [3]. Auto-GPT is powered by self-managing artificial intelligence agents that can self-request commands and execute tasks with minimal human intervention. Self-triggered prompts enable complex research to be conducted without human intervention. Given a user query, Auto-GPT performs step-by-step actions to realize the goal autonomously. In addition to reasoning about the current situation, Auto-GPT can also use past actions to improve decision making [4]. Autonomous AI can learn, think and act without continuous input from humans [5]. In this way, it aims to obtain the result by sequentially generating commands from the given basic commands and their results. While there are evidence suggesting that scaling up Large Language Models (LLM) can lead to certain level of general intelligence, there are still limitations as the lack of long-term memory, limited token length, and the lack of deterministic control over its behaviors, etc. [6].

II. IS AAI POSSIBLE?

The ChatGPT natural language processing model was publicly launched in November 2022 and scaled at an unprecedented rate, reaching 1 million users in five days compared to 300 days for Facebook, 720 days for Twitter, and 75 days for Instagram. The most important factor underlying this rapid spread is its text production and processing feature. ChatGPT can generate high-quality, plausible, human-like written responses to statistical analyses, texts, computer programs, abstracts, or introductions to scientific articles[7]. ChatGPT is able to do this by taking advantage of its extensive data stores and efficient design to understand and interpret user requests and then generate appropriate responses in early natural human language[8]. The use of ChatGPT also presents various potential problems, such as the generation of incorrect or fake information and student plagiarism, and for students it is crucial to introduce them to the limitations of ChatGPT, such as its reliance on biased data, limited current knowledge, and the potential for generating incorrect or fake information [9]. Developed using the GPT-3.5 language model, an improved version of GPT-3, ChatGPT is the result of OpenAI's ongoing work on increasingly secure and useful artificial intelligence systems [10]. In March 2023, shortly after ChatGPT was launched, the GPT 4 is available as final version. GPT-4 builds on the success of its predecessors, GPT-3 [11] and GPT-2 [12]. These models have shown significant improvements in their ability to understand and generate natural languages. With 175 billion parameters, GPT-3 sets new standards for several natural language processing tasks [13]. In the GPT-4 model, the number of parameters reached trillions, significantly increasing its capacity compared to previous models [14]. In addition to human feedback-

Cite (APA) : Fırat, M., Kuleli, S. (2023). What if GPT4 Became Autonomous: The Auto-GPT Project and Use Cases. *Journal of Emerging Computer Technologies*, 3(1), 1-6. Doi: 10.57020/ject.1297961

supported learning, GPT-4 has a rule-based reward model approach compared with GPT-3 [15].

Chat GPT is able to accomplish with its language processing capability, which has been an important milestone in the history of artificial intelligence and made it the top of the agenda in a short time. The use of language models like ChatGPT in education appears as a potential area of interest due to their rich and wide range of applications [16]. Using these models, it will be possible to create personalized and effective learning experiences for individuals at every level of education, in line with each individual's unique learning preferences, abilities, and needs [17]. Going forward, we expect ChatGPT's performance to increase through deep learning, a promising development educators and researchers should pay attention to for its potential applications in teaching and learning [18]. However, it should be noted that ChatGPT's ability to generate highly realistic texts poses a potential risk to the integrity of online exams. Precautions should be taken to prevent this from happening [19,20] LLMs can also be used to develop more natural and sophisticated user interfaces by exploiting their ability to generate contextualized, personalized, and diverse responses to natural language questions asked by users [21]. Chat GPT performs the desired commands thanks to user prompts, that is, user instructions. A prompt is a chat message that the user sends to an LLM, such as ChatGPT, in a chat-based environment [22]. In this case, the most important step in achieving high efficiency from artificial intelligence is to write the correct prompt. With the introduction of AutoGPT, the automatic prompt generation and processing promised by autonomous artificial intelligence has caused users to turn their attention in this direction.

The autonomous transformation of AI has been a topic of discussion since the 1960s [23]. Following the launch of GPT-4 in March 2023, studies on the use of productive artificial intelligence technologies in almost every sector came to the forefront and created a big agenda in the first quarter of 2023. In natural-language processing-based generative artificial intelligence models, directing the right prompt to artificial intelligence is the key to the efficiency of autonomous artificial intelligence. AutoGPT, an open-source Python project, was created to fill an important gap in helping users reach their targeted results by generating automatic prompts from the given commands. Auto-GPT proposes the goal that attempts to make GPT-4 fully autonomous and try to solve the problem that let ChatGPT interact with the internet [24].The emergence of autonomous systems in an increasing number of domains has made it imperative for AI agents to deal with environmental uncertainty through creativity [25]. An autonomous AI application is a system that perceives an environment as part of it and works to influence what it perceives in the future in line with its goals over time. Wang [26] presents basic research on the fundamental theories, discoveries and latest developments in AAI and Symbiotic Human-Machine Intelligence. Autonomous Artificial Intelligence (AAI) is recognized as the general form of AI equivalent to human Natural Intelligence (NI), supported by intelligence science, brain-inspired systems, cognitive computers, intelligent mathematics and systems, humans, and cybernetics. However, AI agents differ from objects in object-oriented computer programs in that they are autonomous, flexible, and have their own control structures.

According to Totschnig [27], "In AI and AI applications, autonomy refers to an artificial entity's ability to act autonomously without human guidance, assuming a fixed goal or utility for evaluating its actions. However, from a philosophical point of view, this concept of autonomy seems unlikely to be implemented, but the rapid proliferation of autonomous AI applications, such as BabyAGI and AutoGPT, which emerged in 2023, proved that the pace of development of artificial intelligence should not be underestimated.

AutoGPT, which performs operations using the GPT-4 model, can work autonomously and perform the steps of thinking, planning, and taking action independently. While working with LLMs, it automatically generates prompts in line with the given command and works until it reaches the result, without the need for users to add any input. Considering that creating the correct prompt and presenting it to artificial intelligence is the most important step in achieving the desired result, AutoGPT will produce effective results by saving users from this step. While there has been growing interest in Auto-GPT-styled agents, questions remain about the effectiveness and flexibility of Auto-GPT in solving real-world decision problems [28].

While recent studies on the use of autonomous artificial intelligence in areas such as health [29], energy production [30], and genetics [31] are available in the literature, there are no studies in the field of open and distance learning. Studying the Chat GPT's ability to enhance autodidactic learning is very important for several reasons. It can identify best practices and approaches for using chatbots and other artificial intelligence (AI) tools in education as well as provide guidance for the future of education and the use of technology in learning [32]. Auto-GPT combines GPT-4's powerful natural language processing capabilities with the ability to auto-author prompts in a single framework, enabling a large number of tasks to be automated with a high level of precision and efficiency. This development shows that LLMs, which could be an important future solution tool for AI, have the potential to radically transform the way we look at business automation. Despite this, studies on Auto-GPT and its potential applications are very limited in the relevant literature. In particular, the experiences of developers using Auto-GPT can provide important insights to explore the potential of this remarkable application. Determining the potential usage areas and limitations of AutoGPT will shed light on future research in this field.

III. RESEARCH PURPOSE

In the first quarter of 2023, studies on the use of productive artificial intelligence technologies in almost every sector came to the forefront and created a significant agenda. The introduction Auto-GPT has raised questions regarding the current use of generative artificial intelligence technologies and in which areas they can be used effectively in the future. This research aims to examine various use cases and experiences of Auto-GPT users. Thus, this study aims to

identify and understand the experiences of Auto-GPT users and to reveal the potential projects and best use cases.

IV. Method

This research was designed based on the case study methodology proposed by Yin [33]. The exploratory case study method was preferred to investigate and interpret the usage areas and experiences of users on the Auto-GPT project, which is available on the GitHub platform, comprehensively and deeply. In the relevant literature, researchers have various definitions and approaches pertaining to the case study [34, 35, 36]. A case study is a method that examines a single situation or event longitudinally, systematically collecting data and examining what happens in real-world settings. Yin [37] defines case study as a research method used when the research focuses on "how" and "why" questions, when the researcher has little or no control over the events, when you study the event or phenomenon within its natural life framework, when the link between the event and real life is not clear enough. According to Yin [38], one of the features that distinguishes a case study from other designs is that it is combined with an existing theory and, depending on its outputs, three types of case studies are defined as exploratory, descriptive, and explanatory. An exploratory case study design is conducted to investigate a new situation, a topic that has not yet been sufficiently researched, and to fill the gap in the literature [39].

In the research, data analyzed were gathered from the discussion thread on the GitHub platform, where users share their experiences and usage scenarios with Auto-GPT. For the purpose of the research the open-ended question posed in the discussion was "What use cases have you found for Auto-GPT?". Sixteen users responded to this open-ended query. The participants provided voluntary consent for the use of their data anonymously. These comments were provided by 16 different users, offering diverse perspectives and potential applications. The responses from the users were subjected to content analysis. Qualitative data took advantage of thematic content analysis, a common method used to identify, analyze, and report patterns or themes in data. The thematic content analysis was conducted in five stages [40]:

1. *Familiarization*: Initially, researchers carefully read through user comments to become familiar with the data and identify patterns or themes at the outset.

2. *Coding*: Each comment was systematically coded by assigning labels to specific sections of the text that represented a particular idea, application, or experience. This process enabled the researchers to effectively organize and categorize the data.

3. *Identifying themes*: The codes were examined and compared. Similar codes were grouped into broader themes. These themes reflect the most important ideas and experiences shared by users.

4. *Review and Refine Themes*: The identified themes were reviewed and refined to ensure that they accurately represented the data. This process aims to verify that the themes are supported by coded data, and that there are clear distinctions between different themes.

5. *Identifying and Naming Themes*: The final step was to create clear and unambiguous descriptions for each theme and associated subthemes to identify and name the themes.

V. Findings

Themes were derived from the codes in the thematic content analysis. Thus, 21 themes were identified. The themes were sorted into four categories based on their common characteristics. The codes, themes, and categories obtained from the content analysis are presented together with the frequencies in Table I.

Table I. Qualitative Findings

Categories	Themes	Code	Frequencies
Use Cases (f=14)	E-commerce	Sales, customer journeys	1
	Ticketing	Customer queries	1
	Data summarization	Open data portal, news articles	2
	Content creation	Cocktail list, Python script, podcast outline	3
	Tool generation	Programs, files	1
	Information organization	Party schools in the US, market research	2
	Investment analysis	Autonomously gather and analyze market data	1
	Product research	Best headphones	1
	Social media management	Generate content, schedule posts, customer inquiries	1
Limitations (f=4)	Token limit	Stuck at limit tokens	1
	Inefficiency	Scraping articles online	1
	Forgetting generated tools	Not using tools effectively	1
	Iteration	Airtight systems, progress after shutdowns	1
Workarounds (f=3)	Data chunking	Bypass token limit	1
	GPT-4 improvements	Better tool utilization	1
	Session continuity	Return to a session after it closes	1
User Experience (f=5)	AI enthusiasm	AI addiction, learning new technologies	1
	Learning programming	Python language	1
	AI agents	Outsource tasks to other AI agents	1
	Autonomous AI capabilities	Conduct research, write reports	2

The theme with the highest frequency among the 21 themes obtained in the study was "Content Creation". This theme is in line with the feature of producing content on different topics, which is considered the most important outputs of use of GPT models. Direct quotes from two user related to this theme are provided below.

"*I tried to let it make "tools" (programs and files) that it would use while on projects. It made a lot and even impressive*

ones i thought. … it would be cool to have it "externalize" and therefore reduce token cost."

"*I used it to come up with a pretty basic cocktail list. All the cocktails and their recipes were quite good...*"

This theme was followed by "*Data Summarization*", "*Information Organization*" and "*Autonomous AI Capabilities*" with two frequencies for each. These three themes are related to the AutoGPT's ability to conduct self-search. When prompts are given in sufficient detail and in a cyclical structure, AutoGPT's independent research feature comes to the fore. Direct quotes from two user's view on data summarization are given below.

"*Search and summarize data works pretty well. I let it write me a list of all data available in the open data portal.*"

"*I'm trying to summarize the news. But still stuck at limit tokens...*"

AutoGPT users emphasized the potential of AutoGPT for searching and summarizing data, but criticised the token limits. About "information organization", a direct quote from a user is given below.

"*I used it to find the top 10 party schools in the US... and then write each schools online application URL to a text file for easier applying. If only I had this 19 years ago when I was applying to college...*"

In this theme, users emphasized the potential of AutoGPT to organize information with the support of AI technology. Finally, here is a direct quote from a user on the topic of "*Autonomous AI Capabilities*".

"*…chunking large data into smaller bits that don't pass the token limit. I believe Auto-GPT does this when scraping articles online but its not really efficient.*"

Here in this theme, users criticize the autonomous AI capabilities with token limits that cause inefficiency. Each of the other 17 themes has one frequency.

By examining the themes in the Workarounds and User Experience categories, it is possible to gain a more detailed understanding of the user experience. There is significant interest and curiosity about AI technology among users. Users are exploring the potential of AI to learn new technologies and expressing their reliance on AI technologies in the process. The results indicate that users are turning to learning programming languages such as Python. This will increase their ability to use these tools more effectively and deepen their understanding of these technologies. Users appreciate Auto-GPT's ability to delegate tasks to other AI agents. This allows users to focus their time and energy on more strategic or creative tasks, while ensuring that routine or repetitive tasks are completed quickly and efficiently. Autonomous AI capabilities, such as conducting research and writing reports, are among the most frequently cited positive aspects by users. This demonstrates AI's ability to respond to user needs and add value.

In addition to the advantages of AutoGPT, some limitations were also mentioned. These were identified as "*token limit*", "i*nefficiency*", "*forgetting generated tools*", and "*iteration*". When analyzing the codes related to these topics, it is observed that the inefficiency and some shortcomings of AutoGPT are emphasized. These limitations can be considered natural considering that AutoGPT is still under development. Participants made suggestions to overcome these limitations. These suggestions included session continuity, removal of the token limit, and further development of the GPT-4 model.

VI. Discussions

The rising trend in AI development has led to discussions of the reduced use of manpower in many fields. Many studies, especially after the introduction of ChatGPT, have revealed important findings regarding the breadth and diversity of its use. AutoGPT, which was created for the autonomous use of the GPT-4 model, can perform many operations such as content generation, access to web content, text summarization, decision-making, and situation analysis with the sequential commands it generates to execute the given command. While there are studies in the literature on the use of AAI agents in areas such as health [41], energy production [42], and genetics [43], it is important to determine the potential for use in education in general, and open and distance learning in particular. AutoGPT's ability to perform operations by minimizing human intervention determines the areas of use in open and distance learning, and reveals the need to draw a framework on how it can be used in this direction.

The high number of students and the diversity of programs in open and distance learning bring about the need for the effective use of AI technologies to save labor and time at many stages of teaching. Content creation, support services, assessment, and evaluation are among the areas that need the most improvement in terms of labor and time [44]. As a result of this research, AutoGPT's ability to perform more operations in fewer steps, to produce content, to summarize the content produced, and to provide guidance in line with the commands entered, reveals that it can be used effectively in these steps of open and distance learning. However, the GPT-4 has some limitations. The first is that it is paid. The fact that GPT-4 is available for a fee as of April 2023 can also be considered one of the limitations in the use of AutoGPT.

VII. Conclusions and Suggestions

This qualitative research is an exploratory case study that aims to examine use-cases and experiences of Auto-GPT users. Data were collected from the GitHub platform. Answers of users to the question of "What are your experiences on the Auto-GPT use-cases*". In response to this question, 16 comments were analyzed and 21 themes were identified. The three important findings of this study are as follows.

- Integration of Auto-GPT into business processes: The results indicate that integrating Auto-GPT into business processes can provide significant productivity gains, especially in data-driven and information management jobs. This means that companies and organizations can use Auto-GPT to optimize their business processes and become more competitive.

- Auto-GPT's limitations and areas for improvement: This research also reveals Auto-GPT's current limitations and areas for improvement. Limitations such as the token limit, inefficiency, forgetting of generated tools, and iteration have

been identified as important areas for the development and improvement of Auto-GPT. Users' suggestions for overcoming these limitations focused on session continuity, removing the token limit, and further development of the GPT-4 model.

- Potential for use in open and distance learning: The results of various use cases and experiences of Auto-GPT show that it can offer significant opportunities for open and distance learning. The effective use of Auto-GPT in areas such as content creation, support services, assessment, and evaluation can contribute to making such learning systems more efficient and user friendly [44].

The results of this study show that Auto-GPT can offer successful use cases in areas such as content creation, data summarization, knowledge organization, and AAI capabilities. In addition, it has been determined that users' experiences such as learning programs, using autonomous applications, conducting research, and writing reports stand out in their experiences. The codes and themes derived from participant opinions about Auto-GPT indicate that this application has high usage potential in open and distance learning, especially in mass teaching systems. However, both the lack of desired maturity in AI technologies and the fact that Auto-GPT is a newly emerged tool that has not yet evolved into a stable application pose challenges for its usage in sensitive areas such as education.

VIII. RECOMMENDATIONS

Continue the development of Auto-GPT and overcome existing limitations. In particular, issues such as token limits and inefficiency must be addressed.

The use of Auto-GPT in the field of education and learning technologies should be encouraged, especially considering that it can be used effectively in open- and distance-learning processes.

Platforms should be created to share experiences and use cases related to Auto-GPT among users, and users should be able to interact on these platforms. In this way, knowledge sharing regarding the potential applications and effective usage methods of Auto-GPT can be increased.

REFERENCES

[1] Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. IEEE transactions on neural networks and learning systems, 32(10), 4291-4308.

[2] Bahja, M. (2020). Natural language processing applications in business. E-Business-higher education and intelligence applications.

[3] Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K. W., Wu, Y. N., ... & Gao, J. (2023). Chameleon: Plug-and-play compositional reasoning with large language models. arXiv preprint arXiv:2304.09842.

[4] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., ... & Sun, M. (2023). Tool learning with foundation models. arXiv preprint arXiv:2304.08354.

[5] Palace, V. M. (2019). What if artificial intelligence wrote this: artificial intelligence and copyright law. Fla. L. Rev., 71, 217.

[6] Yang, H., Yue, S., & He, Y. (2023). Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. arXiv preprint arXiv:2306.02224.

[7] Rahimi, F., & Abadi, A. T. B. (2023). ChatGPT and publication ethics. Archives of medical research, 54(3), 272-274.

[8] Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. Library Hi Tech News, 40(3), 26-29.

[9] Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410.

[10] OpenAI, (2023). Chat GPT. Retrieved from https://openai.com/blog/chatgpt/ on 2 January 2023.

[11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[14] Firat, M. (2023). How chat GPT can transform autodidactic experiences and open education. Department of Distance Education, Open Education Faculty, Anadolu Unive.

[15] Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22312330.v1

[16] Firat, M. (2023). How chat GPT can transform autodidactic experiences and open education. Department of Distance Education, Open Education Faculty, Anadolu University.

[17] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274..

[18] Talan, T., & Kalınkara, Y. (2023). The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course. Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi, 7(1), 33-40.

[19] de Winter, J. C. (2023). Can ChatGPT pass high school exams on English language comprehension. Researchgate. Preprint.

[20] Susnjak, T. (2022). ChatGPT: The end of online exam integrity?. arXiv preprint arXiv:2212.09292.

[21] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. 2023-01-15 A Position Paper. https://www.edu.sot.tum.de/fileadmin/w00bed/hctl/_my_direct_uploads/ChatGPT_for_Good_.pdf

[22] White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. arXiv preprint arXiv:2303.07839.

[23] Radanliev, P., & De Roure, D. (2023). Review of the state of the art in autonomous artificial intelligence. AI and Ethics, 3(2), 497-504.

[24] Guo, T. (2023). Auto Self-Correct ChatGPT. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22688317.v1

[25] Gizzi, E., Nair, L., Chernova, S., & Sinapov, J. (2022). Creative problem solving in artificially intelligent agents: A survey and framework. Journal of Artificial Intelligence Research, 75, 857-911.

[26] Wang, Y. (2023). Autonomous AI (AAI) and Symbiotic Human-Machine Intelligence Systems. In Plenary talk, IEEE International Conference on Systems, Man, and Cybernetics (pp. 3-1).

[27] Totschnig, W. (2020). Fully autonomous AI. Science and Engineering Ethics, 26, 2473-2485.

[28] Yang, H., Yue, S., & He, Y. (2023). Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. arXiv preprint arXiv:2306.02224.

[29] Thodberg, H. H., Thodberg, B., Ahlkvist, J., & Offiah, A. C. (2022). Autonomous artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. Pediatric Radiology, 52(7), 1338-1346.

[30] Noye, S., Martinez, R. M., Carnieletto, L., De Carli, M., & Aguirre, A. C. (2022). A review of advanced ground source heat pump control: Artificial intelligence for autonomous and adaptive control. Renewable and Sustainable Energy Reviews, 153, 111685.

[31] Omelianenko, I. (2022). Autonomous Artificial Intelligent Agents. Machine Learning and the City: Applications in Architecture and Urban Design, 263-285.

[32] Firat, M. (2023). How chat GPT can transform autodidactic experiences and open education. Department of Distance Education, Open Education Faculty, Anadolu Unive.

[33] Yin, R. K. (2009). Case study research: Design and methods (Vol. 5). sage.

[34] Creswell, J. W., & Poth, C. N. (2016). Qualitative inquiry and research design: Choosing among five approaches. Sage publications.

[35] Chmiliar, I. (2010). Multiple-case designs. J. Mills, G. Eurepas & E. Wiebe (Eds.), In: Encyclopedia of case study research (pp. 582-584).

[36] Merriam, S. B. (2013). Nitel araştırma: Desen ve uygulama için bir rehber. Çev.Ed. Sellahattin Turan. Ankara: Nobel Yayın Dağıtım.

[37] Yin, R. K. (1984) Case study research.

[38] Yin, R. K. (2009) Case study research: Design and methods. Vol. 5. Sage.

[39] Swedberg, R. (2020) Exploratory research. The production of knowledge: Enhancing progress in social science, 17-41.

[40] Vaismoradi, M., Hannele T., Terese B. (2013) Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing & health sciences, 15(3), 398-405.

[41] Thodberg, H.H., Thodberg, B., Ahlkvist, J. et al. (2022) Autonomous artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. Pediatr Radiol 52(7), 1338–1346. https://doi.org/10.1007/s00247-022-05295-w.

[42] Noye, S., Martinez, R. M., Carnieletto, L., De Carli, M., & Aguirre, A. C. (2022). A review of advanced ground source heat pump control: Artificial intelligence for autonomous and adaptive control. Renewable and Sustainable Energy Reviews, 153, 111685.

[43] Omelianenko, I. (2022). Autonomous Artificial Intelligent Agents. Machine Learning and the City: Applications in Architecture and Urban Design, 263-285.

[44] Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. Journal of Applied Learning and Teaching, 6(1).

Journal of Emerging Computer Technologies
Research Article
Received: *2023-05-18* | ***Reviewing:*** *2023-05-25* & *2023-09-29*| ***Accepted:*** *2023-09-30* | ***Online:*** *2023-10-01* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1298919

Gamified Violin Playing in Virtual Reality Based Metaverse Environment

Ercan AYKUT
Department of Electrical and Electronic Engineering
Faculty of Engineering-Architecture
Istanbul Gelisim University
İstanbul, Türkiye
eaykut@gelisim.edu.tr
0000-0001-8639-8408
(*Corresponding Author*)

Sena TAŞ
Department of Electronics and Automation
İstanbul Gelişim Vocational School
Istanbul Gelisim University
İstanbul, Türkiye
setas@gelisim.edu.tr
0000-0002-4162-7585

***Abstract*—Violin is one of the most significant musical instruments of both Turkish and western music. It is a stringed instrument played with a bow. Due to its fretless nature, it is relatively difficult to learn and play the instrument, compared with other musical instruments. In this study, a violin has been modeled to be played in virtual reality and metaverse environment is for a better and more entertaining violin education. Even if the student does not actually have a violin in hand, by means of virtual reality glasses, the modeled violin can be seen in the virtual reality environment. The camera of the Oculus Quest 2 Virtual Reality Glasses and haptic gloves are used to sense the both student's finger pressing and bow movements. By means of this study, the student does not need to own a real violin. Considering that a good violin has a high cost, the student will not need to spend high cost to buy the instrument. Besides, the biggest problem of the violin and other stringed musical instruments is the tuning of the instrument. Without a good tuning, the strings will not sound harmoniously and the student will not be able to play correctly because of not hearing the correct sound. Tuning ability needs some experience. But there is no tuning problem in the violin modeled in this study since it does not have real strings which means that even a beginner level student will not have a tuning problem. Also, in this study, while playing various exercises and songs, the musical notes needed to be pressed on the violin are highlighted to ensure that the student confirms whether to play correctly. The study is an alternative material of violin education in a more entertaining way.**

***Keywords*— Metaverse, Virtual Reality, Gamification, Unreal Engine, Violin**

I. INTRODUCTION

Virtual reality (VR) applications were firstly seen in the 1960s and its use has been increasing since then [1]. Investments in virtual reality sector have also been increasing recently. For example, Facebook bought the Oculus brand in 2014 by spending billions of dollars [2]. Furthermore, Facebook changed its name to Meta to adapt to the concept of metaverse [3]. Virtual reality is used in many areas such as education, health, gaming, industry and daily life. There are many studies in the literature on virtual reality applications in different fields.

Han Fang et al. studied wearable haptic technologies for virtual reality and augmented reality (AR). By means of wearable accessories, body movements are perceived as well as hand and finger movements [4]. Adriano da Silva Marinho et al., for the first time, conducted a study on postural stability and cyber diseases in virtual reality game players [5]. Dai-Yun Wu et al. worked on the application of virtual reality in the advertising industry [6]. Jingni Ma et al. studied mental improvement in adults by means of virtual reality [7]. Jing Wen et al. conducted a study in which an electrician explained some experiences to students in a virtual reality environment [8]. Alena Kostyk et al. conducted a study on virtual reality and customer-oriented marketing [9]. Babji Srinivasan et al. conducted a study to enhance chemical safety in laboratories and industry with the use of virtual reality [10]. There are also some studies in the literature on Unreal Engine. Eric Chu et al. suggested an alternative to the visual programming of the Unreal Engine game engine called Blueprints [11].

By means of this study, by using virtual reality glasses, the student can learn to play the violin in an entertaining way without spending high cost for the instrument. Also it allows players to come together online and make music together. The study includes virtual reality, blockchain, NFT and metaverse concepts and contributes to the literature.

II. MATERIAL AND METHODS

A. 3d Design

In this study, firstly, violins and accessories used in the virtual reality environment have been modeled by means of three-dimensional design programs. Software such as Blender, 3d Studio Max, Autodesk Maya can be given as examples of three-dimensional design and animation programs [12]–[14].

Fig. 1. Image of a modeled violin

Later, the created model has been imported to the Unreal Engine game engine as an asset. Many kinds of different assets from different marketplaces can also be used in Unreal Engine. A realistic image has been created by applying appropriate material to these objects in the mesh state. An example image of a modeled violin is shown in Figure 1. The

Cite (APA) : Aykut, E., Taş, S. (2023). Gamified Violin Playing in Virtual Reality Based Metaverse Environment. *Journal of Emerging Computer Technologies*, 3(1), 7-11. Doi: 10.57020/ject.1298919

same process was applied to the other accessories and the materials used in the design have been created.

B. Virtual Reality Glasses

Although most of the virtual reality applications consist of games, applications of education and different sectors are also frequently encountered [15]. Virtual reality has its own rules and it doesn't have to be the same as reality. Thanks to the virtual reality glasses, the user of the glasses immerses into virtual environment [16]. By means of the controllers used with the virtual reality glasses interaction can be added to the game or application. These controllers have the ability of operations such as holding, firing and pressing. Some of the virtual reality glasses used today can be listed as Oculus Quest 2, HTC Vive, HP Reverb G2 and Samsung SM-R323. In this study, Oculus Quest 2 virtual reality glasses have been used to see the modeled violin and accessories in a virtual reality environment. The virtual reality glasses used in the study can be seen in Figure 2.

Fig. 2. Oculus Quest2 Virtual Reality Glasses

C. Sensing The Movements

Body and hand movements of the users should be detected and transferred to the virtual environment in order to add interaction to the virtual reality application or to make it more similar to the real environment. Therefore, carrying, holding, and triggering can be done with the controllers of the virtual reality glasses, while body movements such as hand and finger movements or bending and jumping of the human can be transferred to the virtual environment through wearable technologies. Moreover, the process of touching an object in the virtual reality environment can be felt by the human by means of wearable technology [17].

VR gloves can be assumed as an alternative for VR controllers. By means of the VR gloves, one can feel the immersion at a higher degree. Beside making the controller actions, haptic feedback gloves can also sense the feeling, adding even more immersion into the metaverse. Interaction with virtual objects is possible with hands instead of a controller by wearing VR gloves with haptic feedback technology. It is possible to feel the 3d objects by hands and fingers which is ideal for VR training and simulations, to create an immersive learning experience. When touching an object in 3d environment, haptic gloves provide feedback making vibrations on the fingertips and so it causes to feel the existence of an object [18].

As a sensing element, the built-in camera of the Oculus Quest 2 or the haptic VR gloves seen in Figure 3 are used to detect the user's finger pressing and bow pull and push in this application.

Fig. 3. Haptic feedback gloves

D. Game Engine

The operation of the system is provided by the Unreal Engine 5 game engine. Game engines such as Unity and Unreal Engine are software used to make more beautiful and detailed games in a shorter time [19]. While Unity is programmed with C sharp language, Unreal Engine is programmed with both C++ and a visual programming technique called Blueprints [20]. The demand for the Unreal Engine game engine has been increasing recently, since it does not need a programming language. Therefore, the Unreal Engine game engine has been used in this study. A view from Unreal Engine 5 can be seen in Figure 4.

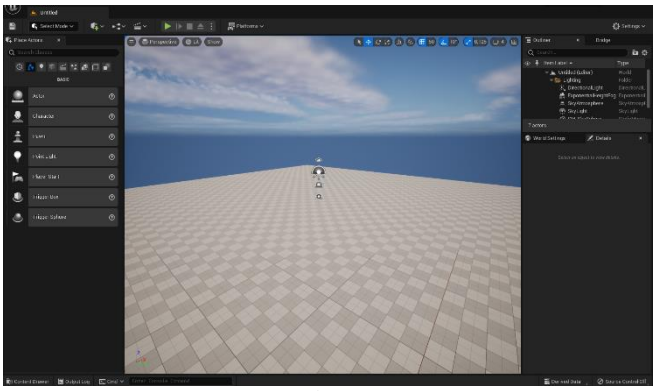


Fig. 4. A view from Unreal Engine

E. Gamification

Gamification is the process of adding game features and visuality to an activity that is not actually a game. Gamification is an activity increasing motivation in education. Especially, pedagogically supported gamification contributes to the versatile development of students [21]. Gamification not only increases cognitive development, but also makes it easier to learn a foreign language [22].

Gamification makes learning more entertaining in order to achieve the desired result in learning. Rewarding in the game environment can provide an interactive environment for the

student. In order to ensure the digitalization of higher education in Turkey, the "YÖK (Council of Higher Education) Digitizing" project was implemented. The Ministry of National Education of the Republic of Türkiye is also switching from traditional systems to digital methods. Therefore, the usage area of gamification is expanding [23]. A view from gamification of the simulation made in the study can be seen in Figure 5.

Fig. 5. A view from Gamification

F. Location to Play

In this study, some locations are modeled and the student can choose the location where to play the violin. This may be a seaside, concert hall or any other place. The student feels to exist in the place of the virtual environment. This may increase the motivation to learn or play the violin. A view from a seaside location can be seen in Figure 6.

Fig. 6. Seaside location

G. Metaverse

The word “metaverse” was firstly used in the Novel Snow Crash written by Neal Stephenson in 1992 [24]–[29]. Even though there is no specific definition, there are recognized technologies and features that are being adopted, such as XR (Extended reality) and blockchain that facilitate the integration and interoperability of decentralized systems to create realistic, immersive virtual worlds and experiences that can be accessed by anyone and any number of concurrent users [30], [31].

Nowadays it is defined as an evolving state medium that can mimic the real world. It is mostly hard to distinguish the human based applications made in the real world and the virtually created world. It is shaped by the specific needs of the users (as engaging in life-like remote collaborations across geographies). It relies upon emerging technologies to support its various requirements (such as verifying a digital object created by a specific person or guaranteeing the success of a financial transaction between two entities) [26], [29]. According to another definition it is described as a medium created by users to be used for other users.

H. NFT

Non-fungible tokens (NFT) are used for recording, verifying and tracking the ownership of the unique assets either physical or digital. Moreover, NFTs can be utilized to represent artworks, futures contracts, music scores, books, and real estate, etc. – any type of item considered as unique or rare. NFTs can be minted, stored and transferred on a blockchain, and hence cannot be obtained by bad intentions [32].

NFTs also indicate the authenticity, transparency and stability of digital work. By means of NFTs having a high degree of confidentiality and transparency, the artworks and other products are not sold in the galleries. Instead, they are sold as NFT and delivered to the user very shortly [33].

Creators can generate both active and passive income by selling virtual assets to consumers who use them for various focuses (such as collaborations, entertainment, data interface, and education). It provides a symbiotic relationship between producer and consumer that fosters the adoption and lasting use of the "metaverse" where creators can earn a living and people can buy digital products, services and experiences. NFTs are a mechanism to facilitate this process via their unique digitalized smart contracts that enable secure transactions based on the blockchain protocol The NFT of the virtual asset is permanently linked to the creator. Additionally, it contains rules of sale to enable both active incomes (direct development of the asset for sale to a user) and passive income.

The main purpose of this study is to learn a musical instrument like the violin, which is difficult to play and learn, with a lower budget and having fun. For this reason, education is divided into levels by gamification. To advance to the next level, the previous level must be completed. When the student wears glasses, the exercise appears on the screen and the student can see the note and its duration also sees where to press on the violin. In addition, the student can customize the color or type of violin. Different colors of violins or accessories can also be bought as NFT (Non Fungible Token) over the Ethereum network.

III. Problems

The violin is a sensitive and difficult musical instrument to learn, because of its fretless and relatively small keyboard [34]. A beginner level violin is inexpensive, but since it is made up of low quality material, it has a short life and it may

have a disturbing sound. Therefore, even as a beginner, the student should buy a good violin, however this requires a relatively high cost. Another difficulty encountered while playing the violin is that the resin applied to the bow causes allergic effects in some people [35]. One of the biggest problems of the violin and all stringed musical instruments is the necessity of tuning. Tuning is the process of adjusting the strings to the required frequency and musical note [36]. The strings of the violin and the required musical note and frequencies are given in Table 1.

TABLE I. NOTE S AND FREQUENCIES OF THE STRINGS OF VIOLIN

	1st String	**2nd String**	**3rd String**	**4th String**
Note name	Sol (G)	Re (D)	La (A)	Mi (E)
Frequency (Hz)	196,00	293,66	440,00	659,26

Since the tuning process requires experience, the student may not tune correctly at the beginning. Even if the tuning is done by the trainer, when the student goes home, then the violin will get out of tune again as a nature of the stringed instruments, and if the student cannot tune, this will cause the string to sound disharmoniously. A tuner is usually used for tuning. The strings are loosened or tightened by the pegs. The tuner indicates whether to loosen or tighten the strings and when the target frequency is achieved, it displays it on the screen. In Figure 7, the picture of a tuner is presented. If there is no good tuning, the musical instrument will not sound properly and the student cannot learn correctly or music cannot be made together with other musical instruments in harmony which may cause the student to give up the musical instrument and stop learning it. One of the real-life problems is that when a student goes home after the class, the student may be unsure of the accuracy of playing, during practicing or doing homework

Fig. 7. A view from Gamification

Since there is no trainer at home, the student may not be sure about playing correctly. Another problem in real life is that people may not come together and therefore may not make music together in case of pandemics and epidemics.

This study was made to be a solution to the problems mentioned above.

IV. DISCUSSIONS

In this study, a virtual reality application is developed aiming to popularize and support violin education and to play it in a funny and more effective way under Unreal Engine 5 platform. The project can also be assumed as a simulator which is just for entertainment but will support real life education. One of the advantages of the study is that there is no need to buy a musical instrument at high cost, since a real violin is not needed. Similarly, there is no need to apply real resin since real bow is not used. Therefore, thanks to this study, the allergic effect of the resin will also be prevented. Another contribution of the study is that it eliminates the tuning problem. Because the modelled violin does not have real strings, the sounds are at the exact frequency and there is no tuning problem.

Thanks to this study, education has been made fun by using the Unreal Engine 5 game engine and gamification. In order to popularize the education, a model called Play2Earn was applied. As the student progresses, they earn coins and can shop with them in the system. Accessories are sold as NFT. The student can buy violins or accessories in any color and feature. Blockchain and Ehereum infrastructure is used for shopping.

Another advantage of the study is that when putting on the virtual reality glasses, the location where should be pressed on the violin and where the next note corresponds on the violin can be visualized. Therefore, the student is sure about whether to sound correct pitch.

In addition, in this study, several people can make music together at the same time. It allows musicians who cannot come together due to the pandemic and epidemic disease or because they are far away, to make music together.

This study is not a replacement for real life violin education since no methods such as Suzuki method [37] or Ömer Can method [38] which are the most famous violin education methods in Turkey, are followed. In future work, the education can be made by following the methods and it may become an alternative education.

V. CONCLUSIONS

This study will contribute to the literature, especially in the field of virtual reality applications. Those who want to learn the violin will learn to play the violin lovingly and willingly in a more fun and gamified way. Thanks to this study, the high cost of the musical instrument can be saved, and the absence of a tuning problem is another advantage of this study. Coins earned as you level up encourage continuous leveling up. By means of gamification, student wants to learn and play more willingly. This study also integrates virtual reality and haptic feedback wearables. In addition, thanks to this study, distant friends can come together in a metaverse environment and make music together through the same system. The shortcoming of this study is the lack of fine details of playing techniques such as legato and staccato. As the next study, more efficient and entertaining educational material will be created by applying a similar system to other musical instruments. It is aimed to increase the quality of education with new glasses and wearable technologies.

Although there are many applications of virtual reality in education sector, due to the developing technology, more detailed applications with more feeling and immersion will be achieved. By means of gamification, a better education can be supplied together with haptic feedback virtual reality applications.

REFERENCES

[1] Sutherland, I. E. (1965). The ultimate display. In Proceedings of the IFIP Congress, Vol. 2, pp. 506-508.

[2] Syamimi, A., Gong, Y., & Liew, R. (2020). VR industrial applications—A singapore perspective. Virtual Reality & Intelligent Hardware, 2(5), 409-420. Doi: 10.1016/j.vrih.2020.06.001.

[3] Kocabay-Şener, N. (2021). Facebook Nasıl "Meta" laştı?. Yeni Medya, 2021(11), 174-179.

[4] Fang, H., Guo, J., & Wu, H. (2022). Wearable triboelectric devices for haptic perception and VR/AR applications. Nano Energy, 96, 107112., doi: 10.1016/j.nanoen.2022.107112.

[5] Ada Silva Marinho, A., Terton, U., & Jones, C. M. (2022). Cybersickness and postural stability of first time VR users playing VR videogames. Applied Ergonomics, 101, 103698., doi: 10.1016/j.apergo.2022.103698.

[6] Wu, D. Y., Lin, J. H. T., & Bowman, N. D. (2022). Watching VR advertising together: How 3D animated agents influence audience responses and enjoyment to VR advertising. Computers in Human Behavior, 133, 107255., doi: 10.1016/j.chb.2022.107255.

[7] Ma, J., Zhao, D., Xu, N., & Yang, J. (2023). The effectiveness of immersive virtual reality (VR) based mindfulness training on improvement mental-health in adults: A narrative systematic review. Explore, 19(3), 310-318, doi: 10.1016/j.explore.2022.08.001.

[8] Wen, J., & Gheisari, M. (2021). VR-Electricians: Immersive storytelling for attracting students to the electrical construction industry. Advanced Engineering Informatics, 50, 101411, doi: 10.1016/j.aei.2021.101411.

[9] Kostyk, A., & Sheng, J. (2023). VR in customer-centered marketing: Purpose-driven design. Business Horizons, 66(2), 225-236. doi: 10.1016/j.bushor.2022.06.005.

[10] Srinivasan, B., Iqbal, M. U., Shahab, M. A., & Srinivasan, R. (2022). Review of virtual reality (VR) applications to enhance chemical safety: from students to plant operators. ACS Chemical Health & Safety, 29(3), 246-262. doi: 10.1021/acs.chas.2c00006.

[11] Chu, E., & Zaman, L. (2021). Exploring alternatives with unreal engine's blueprints visual scripting system. Entertainment Computing, 36, 100388. doi: 10.1016/j.entcom.2020.100388.

[12] Hendriyani, Y., & Amrizal, V. A. (2019, November). The Comparison Between 3D Studio Max and Blender Based on Software Qualities. In Journal of Physics: Conference Series (Vol. 1387, No. 1, p. 012030). IOP Publishing. doi: 10.1088/1742-6596/1387/1/012030.

[13] Stojanov, I., Ristevski, B., Kotevski, Z., & Savoska, S. (2016). Application of 3ds Max for 3D Modelling and Rendering, in Proccedings of the ICAIIT2016, 133–144. doi: 10.20544/AIIT2016.17.

[14] Wood, A. (2014). Behind the scenes: A study of Autodesk Maya. Animation, 9(3), 317-332. doi: 10.1177/1746847714546247.

[15] Çankaya, S. (2019). Use of VR headsets in education: A systematic review study. Journal of Educational Technology and Online Learning, 2(1), 74-88. doi: 10.31681/jetol.518275.

[16] Bayraktar, E., & Kaleli, F. (2007). Sanal gerceklik ve uygulama alanlari. Akademik Bilişim'07 - IX. Akademik Bilişim Konferansı, Dumlupinar University, Kütahya. 1(6).pp. 253–259.

[17] Stasolla, F. (2021). Virtual reality and wearable technologies to support adaptive responding of children and adolescents with neurodevelopmental disorders: A critical comment and new perspectives. Frontiers in Psychology, 12, 720626., doi: 10.3389/fpsyg.2021.720626.

[18] Perret, J., & Vander Poorten, E. (2018, June). Touching virtual reality: a review of haptic gloves. In ACTUATOR 2018; 16th International Conference on New Actuators, Bremen, 270–274.

[19] Christopoulou, E., & Xinogalos, S. (2017). Overview and comparative analysis of game engines for desktop and mobile devices., International Journal of Serious Games,4(4), doi: 10.17083/ijsg.v4i4.194.

[20] Juliani, A., Berges, V. P., Teng, E., Cohen, A., Harper, J., Elion, C., ... & Lange, D. (2018). Unity: A general platform for intelligent agents. arXiv preprint arXiv:1809.02627. Accessed: Feb. 13, 2023. [Online]. Available: http://arxiv.org/abs/1809.02627

[21] Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. The International Journal of Information and Learning Technology, 35(1), 56-79. doi: 10.1108/IJILT-02-2017-0009.

[22] Öden, M. S., Bolat, Y. İ., & Göksu, İ. (2021). Kahoot! as a gamification tool in vocational education: More positive attitude, motivation and less anxiety in EFL. Journal of Computer and Education Research, 9(18), 682-701., doi: 10.18009/jcer.924882.

[23] Karoğlu, A. K., Kübra, B. A. L., & Çimşir, E. (2020). Toplum 5.0 sürecinde Türkiye'de eğitimde dijital dönüşüm. Üniversite Araştırmaları Dergisi, 3(3), 147-158. doi: 10.26701/uad.815428.

[24] Mystakidis, S. (2022). Metaverse. Encyclopedia, 2 (1), 486–497. doi: 10.3390/encyclopedia2010031.

[25] Narin, N. G. (2021). A content analysis of the metaverse articles. Journal of Metaverse, 1(1), 17-24.

[26] Park, S. M., & Kim, Y. G. (2022). A metaverse: Taxonomy, components, applications, and open challenges. IEEE access, 10, 4209-4251. doi: 10.1109/ACCESS.2021.3140175.

[27] Rameshwar, J. R. (2022). Caribbean metaverse development: A literature review perspective. Journal of Metaverse, 2(2), 83-99. doi: 10.57019/jmv.1120470.

[28] Rehm, S. V., Goel, L., & Crespi, M. (2015). The metaverse as mediator between technology, trends, and the digital transformation of society and business. Journal For Virtual Worlds Research, 8(2), 1-6. doi: 10.4101/jvwr.v8i2.7149.

[29] Zhao, Y., Jiang, J., Chen, Y., Liu, R., Yang, Y., Xue, X., & Chen, S. (2022). Metaverse: Perspectives from graphics, interactions and visualization. Visual Informatics, 6(1), 56-67., doi: 10.1016/j.visinf.2022.03.002.

[30] Damar, M. (2021). Metaverse shape of your life for future: A bibliometric snapshot. Journal of Metaverse, 1(1), 1-8.

[31] Dionisio, J. D. N., Iii, W. G. B., & Gilbert, R. (2013). 3D virtual worlds and the metaverse: Current status and future possibilities. ACM Computing Surveys (CSUR), 45(3), 1-38. doi: 10.1145/2480741.2480751.

[32] Ante, L. (2022). The non-fungible token (NFT) market and its relationship with Bitcoin and Ethereum. FinTech, 1(3), 216-224., doi: 10.3390/fintech1030017.

[33] Oduncu, S. (2022). NFT, Kripto Sanatı ve Türkiye'deki Yansımaları. Art-e Sanat Dergisi, 15(29), 195-224. doi: 10.21602/sduarte.1080813.

[34] Kesendere, Y., Sakin, A. Ş., & Acar, A. K. (2020). Educators' views on online/distance violin education at Covid-19 outbreak term. Journal for the Interdisciplinary Art and Education, 1(1), 1-19. doi: 10.29228/jiae.1.

[35] Aytekin, A., Toğral, A., & YILMAZ, Ö. (2015). Epoksi Reçineleri ve Mesleki Dermatozlarla İlişkisi. Ankara Medical Journal, 15(2), 96-101. doi: 10.17098/amj.86443.

[36] Sturman, R. (2012). A dynamical systems approach to musical tuning. Dynamical Systems, 27(1), 131-143. doi: 10.1080/14689367.2012.656300.

[37] Akutsu, T. (2020). Changes after Suzuki: A retrospective analysis and review of contemporary issues regarding the Suzuki Method in Japan. International Journal of Music Education, 38(1), 18-35. doi: 10.1177/0255761419859628.

[38] Başak, N. B., Bağcı, H., & Can, Ü. K. (2020). Türkiye'de Klasik Batı Müziği Alanında Yayınlanmış Keman İle İlgili Öğretici Ulusal Kitapların İncelenmesi. Motif Akademi Halkbilimi Dergisi, 13(30), 917-935., doi: 10.12981/mahder.679746.

Journal of Emerging Computer Technologies
Research Article
Received: *2023-10-17* | ***Reviewing:*** *2023-10-19* & *2024-01-07*| ***Accepted:*** *2024-01-08* | ***Online:*** *2024-01-08* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1375605

A Dynamic Method and Program for Disease-Based Genetic Classification of Individuals

Onur Çakırgöz
Department of Computer Engineering
Bartın University
Bartın, Turkey
onurcakirgoz@bartin.edu.tr
0000-0002-9347-1105

Süleyman Sevinç
Labenko Bilişim A.Ş.
İzmir, Turkey
suleysevinc@gmail.com
0000-0001-9052-5836

***Abstract*—Personalized medicine is gaining increasing importance. However, genetic-based diseases have different underlying genetic factors, requiring separate relative risk models for each disease. In addition to these difficulties, comparing individuals according to their genetic characteristics and determining a personalized treatment method based on this, is a separate problem which is very difficult to do manually. In this study, a dynamic classification method and program is proposed for disease-based classification of individuals according to their genetic characteristics. To the best of our knowledge, this is the first generic method which performs disease-based classification of individuals. In the developed program, relative risk models containing only genetic factors are an input of the program and a common format has been created for this purpose. Our generic classification method classifies people by using information from any relative risk model rearranged according to the common format. Thanks to this program, relative risk models can be managed from a single point, many people can be classified based on their genetic characteristics, and individuals, who are genetically most similar to any person, can be determined by experts using the outputs (relevant tables) of the program.**

Keywords—personalized medicine, computational medicine, genetic classification, relative risk model, genetic similarity, genetic variation

I. INTRODUCTION

Many common human diseases and traits are affected by several genetic and environmental factors [1, 2]. To investigate the genetic variants contributing to these human diseases, researchers do candidate gene studies and Genome-wide association studies (GWAS) [3-5]. Until today, GWAS and gene studies have determined many variants associated with diseases and have provided so many relative risk models [3, 6]. As known, there are so many genetic diseases, and these diseases can be classified as single gene mendelian diseases and complex diseases [7, 8]. In single gene mendelian diseases, as is evident from its name, the variant or variants associated with the disease present at one gene. Additionally, the number of variants associated with the disease is relatively too small [7]. Clinicians can easily diagnose the single gene mendelian diseases most of the time, but unfortunately the things they can do medically are limited [9]. On the other hand, in complex diseases, there might be tens of disease-associated variants [10-12]. For instance, the number of genetic variants contributing to the Crohn's disease is 32 [12]. On the other hand, the number of variants associated with the diseases such as age-related macular degeneration [13], type 2 diabetes [14, 15], early onset myocardial infarction [16] are 5, 18 and 9, respectively. Contrary to the single gene mendelian diseases, clinicians often have more opportunities in complex diseases [4, 17].

There are many genetic-based complex diseases and the numbers of variations associated with these diseases vary considerably. Beyond the number of variations, the properties of the variations vary as well [3, 4, 8]. Most of the variations associated with the diseases increase the risk of developing the disease, whereas others decrease the risk (protective). Sometimes reference allele has high risk; and sometimes alternate allele has high risk. On the other hand, the genetic characteristics of the diseases or traits might differ from population to population or from region to region [18]. A remarkable information is that the vast majority of GWAS and other genetic studies have been limited to European ancestry populations [3,7, 19, 20]. Fewer studies have been carried out in non-European countries (especially populations of under-developed or developing countries) and these studies have determined intriguing new variants. Due to the economic reasons, however, the physicians serving in under-developed or developing countries might have to use the relative risk model generated for a different population (at least until a genome-wide association study is made in its own population for that disease).

With the understanding of the effect of genetic factors, personalized medicine concept has entered our lives [21-24]. Unlike traditional medicine, personalized medicine has adopted the person-specific treatment approach [21, 22]. At this point, the most important mainstay of personalized medicine is genetic characteristics and genetic variations [24, 25]. Since a substantial portion of individual differences in the predisposition to complex disease is due to genetic variants, identifying these variants provides better prevention, diagnosis, and treatment of disease [26]. Within the scope of personalized medicine approach, physicians can inform individuals as to how they can behave to prevent the disease [24]. Furthermore, the individuals who caught a disease and have undergone a successful treatment process, their personal genotypes and the applied treatment method can be stored in a database. When a new patient who has contracted the same disease is admitted to the clinic, clinicians can use the system to identify patients who are genetically most similar to that patient. They can then apply a previously successful treatment method. Unfortunately, to perform these medical applications, a dynamic program is needed that can automatically classify individuals based on their genetic characteristics (disease-

Cite (APA) : Çakırgöz, O., Sevinç, S. (2023). A Dynamic Method and Program for Disease-Based Genetic Classification of Individuals. *Journal of Emerging Computer Technologies*, 3(1), 12-20. Doi: 10.57020/ject.1375605

based) and support relative risk models. Based on this, in this study, a disease-based genetic classification method was designed. This method supports all types of genetic variations. Additionally, a dynamic desktop application was developed. This application utilizes the aforementioned classification method and also supports relative risk models.

In bioinformatics field, after the Human Genome Project [27, 28], the International HapMap Project [29] and 1000 Genome Project [30-33] are considered important turning points. These projects, which catalog the human genetic variations, are put in a separate place. The variation data of 2504 anonymous people were published by 1000 Genomes Project as VCF [34] and BCF [35] files. Within this study, the personal variation data of 2504 people published by 1000 Genomes Project were used.

The sections of the article are as follows: In the second section, firstly, the disease-based genetic similarity approach is mentioned, and the basic classification method is introduced, then, the common format created for the relative risk models is explained, and finally, the developed dynamic program and the components that make up this program are explained. In the third chapter, after loading the sample data to the program, the results produced by the program are shown and these results are discussed. Final remarks are given in the last section.

II. MATERIALS AND METHODS

A. Disease-Based Genetic Similarity and The Basic Classification Method

To be able to perform personalized medicine applications, finding the genetic similarities and dissimilarities of individuals between each other is sometimes required. Considering that, human DNA consists of 3.2 billion base-pairs, humans are diploid and millions of variations exist in the genome of each person, it is seen that the concept of genetic similarity is actually a very complex and broad subject; but what is meant here is disease-based genetic similarity (or classification). As is known, genome-wide association studies aim to identify the variations that influence the disease and to what extent they affect it, by evaluating a number of variations. Thereby, since one of the main objectives is to find genetically most similar individuals to an individual who caught a particular disease, numerous variations which are not associated with that disease should not be considered in genetic similarity between individuals. Besides, since the variations associated with each disease differ, genetic similarity between individuals must be found separately for each disease. As a result, relative risk models generated by genome-wide association studies should be used as the basis for disease-based classification. In parallel, the risk estimations and groups (genotype combinations) specified in the relative risk model are used in our classification method.

The genetic elements that reveal diseases are variations (single nucleotide polymorphisms, short insertions and deletions, and structural variants) occurring in specific locations. At the level of phenotype, another important factor which may determine the emergence of the disease is the presence or absence of genetic variations in both alleles of the person. On the other hand, as we know, the allele counts (reference allele + alternate alleles) of the variations vary. For example, while any variation may have only 2 alleles, another variation may have 3-4 alleles. Given that the human genome is diploid, the number of possible genotypes naturally varies. Depending on the number of alleles of the variation, the number of genotypes that can form is determined according to (1). Thus, when we want to classify individuals based on a specific variation that leads to disease, we can obtain the class count through (1).

$$G_{Nd} = A_{Nd}{}^{2} - \binom{A_{Nd}}{2} \quad (1)$$

The symbols used in this equation and in subsequent equations are shown in Table 1.

TABLE I. SYMBOL TABLE FOR THE DISEASE-BASED CLASSIFICATION APPROACH

Symbol	Explanation
i	Class number
K_d	In how many steps the genotypes of the d^{th} variation vary, that is, the step count of the d^{th} variation
G_{Nd}	Number of genotypes of d'th variation associated with the disease
A_{Nd}	Number of alleles of d'th variation associated with the disease
AV_N	Number of variations associated with the disease
G_{CCd}	The column corresponding to the genotype – Numerical equivalent of the genotype regarding the d^{th} variation
A_{LI}	The index of the lower index allele constituting the genotype
A_{HI}	The index of the higher index allele constituting the genotype
PC_N	Number of possible classes associated with the disease
N_{Nd}	Number of nodes at depth d of the classification tree
UN_{Nd}	Number of unnecessary nodes at depth d of the classification tree
UN_{TN}	Total number of unnecessary nodes on the classification tree

In fact, the number of genotypes that can occur is the square of the number of alleles. But, since we consider genotypes from a functional standpoint and (most of the time) half of the heterozygous genotypes are functionally equivalent to the other half, we subtract half of the heterozygous genotypes from the square of the number of alleles. (If heterozygous genotypes are not functionally equivalent, (1) should be updated accordingly.) For example, suppose that the number of alleles of a variation is two and these are 'A' (reference) and 'T' (alternate and haplotype increasing disease risk) bases. In this case, the genotypes that can be formed are "AA", "TT", "AT" and "TA". Since "AT" and "TA" are functionally equivalent, both are accepted as one. The resulting classes can be defined as: the variation does not exist in both alleles ("AA", class-1), exists only in one allele ("AT" or "TA", class-2), and exist in both alleles ("TT", class-3).

Genetic similarity relationship between classes is a process beyond that and calculating genetic similarities between classes is beyond the scope of this study. Describing and determining the genetic similarity relationship clearly between the resulting genotypes (classes) is not easy most of the time. Since the ultimate goal of doctors is to apply a previously successful personalized treatment to a similar patient, different parameters such as drug-protein interaction may also come into play in the genetic similarity relationship here. In this respect, determining the disease-based genetic similarity relationship is a process that should be done after classification, sometimes it depends on different parameters and therefore, the main decision maker is the doctors (experts). The generic classification method and program proposed in this study systematically presents the relevant classification results to the users. If it is assumed that there are no different parameters in the genetic similarity relationship between classes in the above example, the similarity relationship can be expressed as: The class which is genetically most similar to both classes 1 and 3, is class 2. To find the class, which is genetically most similar to class 2, it is required to look at the risk estimations specified in the relative risk model of the disease. For instance, assume that the classes' disease risks are specified as 1, 3, and 9 in the model, respectively; in that case, due to the lower risk difference, class 1 is the genetically most similar class to class 2. Therefore, sometimes the relative risk model can provide insight into the disease-based genetic similarity relationship.

Genome-wide association studies have shown that the number of susceptibility variants associated with any disease is more than one most of the time. Therefore, in these cases, classification depicted in the previous example is not sufficient. In the case that the number of susceptibility variants associated with the disease is more than one, the resulting number of classes can be obtained using (2). In our method, each different combination of genotypes corresponds to a different class.

$$\boldsymbol{PC_N = \prod_{d=0}^{AV_N-1} G_{Nd}} \tag{2}$$

When viewed from the perspective of computer science, the image that emerges from the classification approach of humans depending on the variations associated with the disease or treat is a tree. In other words, for the classification of individuals based on their genetic characteristics, tree is the first data structure which comes to mind. Of course, this tree data structure would be specific to the problem of classification of people according to their genetic characteristics. As a result, the nodes of this tree are variations, and the branches are genotypes. On the other hand, any path on the tree corresponds to a class (the genotype characteristics of the class), and the individuals involved in a class can be thought as a leaf of the tree. The tree method might seem like an efficient way at first glance, but all the nodes at the same depth have the same value (variation). This means that this tree uses redundant nodes. In addition, the number of unnecessary nodes in each level increases exponentially, depending on the depth of that level. Equation (3) is used to find the number of nodes at any depth of the disease-based classification tree.

$$\boldsymbol{N_{Nd} = \begin{cases} 1, & d == 0 \\ \prod_{k=0}^{d-1} G_{Nk}, & d > 0 \end{cases}} \tag{3}$$

As we have already stated, the values of all nodes at a particular depth of this tree are the same. Actually, only one of these nodes is enough. From this point of view, we can use (4) to calculate the number of unnecessary nodes at a particular depth.

$$\boldsymbol{UN_{Nd} = N_{Nd} - 1} \tag{4}$$

Equation (4) calculates the number of redundant nodes only at a certain depth of the tree. If we want to calculate the number of all unnecessary nodes in the tree, the solution is quite simple. For this, it is necessary to sum the number of unnecessary nodes in each level of the tree. Accordingly, Equation (5) was developed to acquire the total number of redundant nodes on the tree.

$$\boldsymbol{UN_{TN} = \sum_{d=0}^{AV_N-1} UN_{Nd}} \tag{5}$$

The situation that the disease-based classification tree uses redundant nodes and the high cost of the search operation required to find out which class an individual belongs to, have directed us to another data structure. At the beginning, we noticed that the nodes (variations) and the branches (genotypes) of the tree can be stored in a table. In this way, all the variations are stored only once in the table. For instance, when the number of susceptibility variations associated with the disease is 2, the resulting table is seen in Table 2. In this example, the number of alleles (haplotypes) of both variations is two and the respective alleles are indicated in the table. As seen from the table, the number of classes for this disease is found as 9 according to (2), and these classes are numbered from 0 to 8. On the two lines below the class numbers, the corresponding genotypes to that class appear.

TABLE II. CLASSES AND THEIR GENOTYPE PROPERTIES IN THE CASE THAT $AV_N = 2$

Variations	**Classes and Corresponding Genotypes**								
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
d=0 (Alleles=G,C)	G|G	G|G	G|G	G|C	G|C	G|C	C|C	C|C	C|C
d=1 (Alleles=T,A)	T|T	T|A	A|A	T|T	T|A	A|A	T|T	T|A	A|A

The data in Table 2 can be stored in a two-dimensional array. As a matter of fact, our first thought was in this direction as well. Afterwards, while doing operations on the table, we realized that the genotype combinations actually form a pattern, a connection can be established between the classes and the genotypes by using this pattern, and therefore, it is unnecessary to keep the data in Table 2 in any data structure. As can be seen from Table 2, classes are represented by numbers. On the other hand, genotypes are represented by nucleotide bases (letters). In addition, since the alleles of each variation vary, naturally, genotypes also vary. In fact, the important point here is the placement of the genotypes on the table. The genotypes are located in a certain order on the table. For instance, the genotypes of the first variation vary in every 3 steps. On the other hand, this number is 1 for the second variation, and when the genotypes of both variations are

considered together, the resulting genotype combinations form a pattern. Thanks to the pattern that the genotype combinations form, a link can be established between the classes and the genotypes, but the difficulty here is that the genotypes are represented by nucleotide bases. The formulas that will establish the link between genotypes and classes need numerical data. Therefore, primarily, we need to convert the genotypes into numbers. The numerical equivalents of the genotypes in Table 2 are shown in Table 3.

TABLE III. THE NUMERICAL EQUIVALENTS OF THE GENOTYPES IN TABLE 2

Variations	**Classes and Corresponding Genotypes**								
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
d=0 (Alleles=G,C)	0	0	0	1	1	1	2	2	2
d=1 (Alleles=T,A)	0	1	2	0	1	2	0	1	2

Regardless of the number of alleles of the variation, the genotypes that are formed need to be converted into numbers automatically. In other words, there is a need for an equation for the process of converting the genotype to number. For this purpose, (6) was developed. At this point, we must specify that the alleles constituting the genotype are located on the respective "Alleles" array and the A_{LI} and A_{HI} variables in (6) take the values of the indices of these alleles. "Alleles" arrays are the important attributes found in the Variant array, and the Variant array will be described in the section "The Dynamic Program and Its Components".

$$\boldsymbol{G_{CCd} = \left(\sum_{j=0}^{A_{LI}} (A_{Nd} - j)\right) - (A_{Nd} - A_{HI})} \quad (6)$$

The following "Convert_into_genotype(index)" function has been developed to transform the integer back to the genotypes. The "Convert_into_genotype(index)" function is the inverse of (6). Since it is not possible to express this transformation in the form of equation, only the pseudo-code of the function is given here. The variable named "A" in this function represents the "Alleles" attribute explained in the previous paragraph.

```
Convert_into_genotype(index)
 1 Let alleles be a string array of length 2
 2 Let temp and temp2 be integers
 3 temp = index;
 4 for(i = 0; i < A.Length; i++)
 5     temp2 = temp;
 6     temp -= (A.Length-i);
 7     if (temp >= 0)
 8         continue;
 9     else
10         alleles[0] = A[i];
11         alleles[1] = A[i + (temp2 % (A.Length - i))];
12         break;
13 return alleles;
```

In the above section, mainly the conversion of the genotype to the number and the conversion of the number to the genotype were described; namely, the processes and equations related to a single genotype were mentioned. From this point on, the formulas necessary for the relation between the genotypes (number equivalent) and the classes will be addressed. First of all, the relation between the classes and the genotypes is bidirectional. Namely, the genotypes of any class whose number is specified can be determined, or vice versa. Equation (7), which was developed for the first direction of the relation, is below. Thanks to this equation, the genotype (numerical equivalent) corresponding to the respective level (d^{th} variation) of any class, whose number ("i") is specified can be easily found. The "d" parameter of this formula represents the order of the respective variation.

$$\boldsymbol{f(i, d, AV_N) = \lfloor (i/K_d) \rfloor \,\%\, G_{Nd}} \quad (7)$$

The "K_d" variable utilized in the above formula and how this variable is calculated are not yet disclosed. This variable indicates in how many steps the genotypes of the d^{th} variation vary, that is, the step count of the d^{th} variation. For instance, if we look at Table 3, the length of steps in the 0^{th} variation is 3, that is, the genotypes change in every three steps. On the other hand, this value is 1 for the first variation. Equation (8) developed to compute the "K_d" value is given below. There are two different parts in this formula. In the upper section, the step count of the variation with the greatest level is determined. As can be seen from the formula, this value is always 1. In the bottom section, the step counts of other variations are calculated. Here, in order to find the step count of the variation at level d, the genotype numbers of the variations in the upper levels are multiplied, starting from the $(d+1)^{th}$ level.

$$\boldsymbol{K_d = \begin{cases} 1\,, & d == AV_N - 1 \\ \prod_{k=d+1}^{AV_N - 1} G_{Nk}\,, & d < AV_N - 1 \end{cases}} \quad (8)$$

Equation (7), as stated above, is used to find the genotype (numerical equivalent) of a given class for only one variation. If we want to calculate all the genotypes corresponding to a given class, then we must use (7) as the number of variations. The algorithm that accomplishes this process is given below. The "Convert_to_Genotypes()" algorithm has two parameters and these are the respective symbols of (7). Accordingly, the parameter "N_V" represents the number of variations, which is denoted with "AV_N" in (7). The line 3 of the algorithm corresponds to (7). Via the for loop in line 2, all the genotypes (numerical equivalents) corresponding to a given class are obtained uncomplicatedly by executing (7) as the number of variations. In the following algorithm, if we want to obtain genotypes instead of numerical equivalents, we need to convert the type of the "genotypes" variable into a two-dimensional string array and call the "Convert_into_genotype(index)" method on line 4 for "g".

```
Convert_to_Genotypes(int class_no, int N_V)
 1 genotypes is an integer array of length N_V
 2 for (int d = 0; d < N_V; d++)
 3     int g = (Math.Floor(class_no / Kd) % GNd)
 4     genotypes[d] = g;
 5 return genotypes;
```

We have stated before that the relation between the classes and the genotypes is bidirectional. Also, in the above section, we have shown the formulas and algorithms required to determine the genotypes of any class whose number is specified. The formulas and algorithms developed to create the second direction of the relation between classes and genotypes

will be described in this section. Equation (9) was developed to calculate the number of any class whose genotypes are given, in other words, to find the class to which any person belongs. Based on the genotypes of an individual, his/her class can be computed readily by using (9).

$$i = \sum_{d=0}^{AV_N - 1} G_{CCd} \times K_d \quad (9)$$

The algorithm Convert_to_Class_Number(), which corresponds to (9), is below. This algorithm takes only one parameter: a string array that holds the genotypes of the individual. In the fourth line of the algorithm, the numerical equivalent of the d^{th} genotype is calculated. On the other hand, the step count of the d^{th} variation is determined in the fifth line of the algorithm. Note that the calculation of the respective values ("G_{CCd}" and "K_d") in these two lines is not explicitly stated.

```
Convert_to_Class_Number(string[] genotypes)
1 int number1 = 0, number2 = 0;
2 int class_number = 0;
3 for (int d = 0; d < genotypes.Length; d++)
4     number1 = Compute GCCd
5     number2 = Compute Kd
6     class_number += (number1 * number2);
7 return class_number;
```

Individuals, who are genetically (disease-based) most similar to a given individual, are those that are involved in the same class with that given individual. Accordingly, the first place that must be looked is the class that individual belongs to. If there are no other individuals who are involved in the class of a given individual, other classes should be searched. At this point, experts can identify genetically most similar people to a person by looking at the relevant tables that are the outputs of the developed program.

B. Common Format for Relative Risk Models

One of the ordinary works done by the physicians working in the department of medical genetics is calculating the disease risks of individuals. In addition, clinicians who work in the departments such as heart, internal medicine and oncology search previous patients who are genetically most similar to a new patient and apply a successful treatment method previously applied. But, unfortunately, clinicians perform such tasks manually. This situation both leads to a waste of time and may cause confusion. Reaching all the relative risk models through a single application will be a significant convenience for clinicians/physicians. For that, before the development of the program, a common format was created for relative risk models. According to this approach, relative risk models are the inputs of the program and are independent from the program. Alterations made in any relative risk model or developing a novel relative risk model will not affect the program in any way. Thus, clinicians will be able to choose the relative risk model of any disease in an easy way through the program.

The allelic architecture (number, effect size, reference base, alternate base and frequency of susceptibility variants) differs across diseases. For instance, the number of identified variants is 5 for age-related macular degeneration (AMD), which is a common disease, whereas the number of identified variants is 32 for Crohn's disease. Apart from this, most variants identified confer increments at risk, whereas the remaining ones confer decrements at risk. Also, scientific research groups can sometimes handle the rare variants, which increase or decrease the disease risk at similar rates in a single category, in order to simplify the relative risk models as much as possible. During the development of the common format for the relative risk models, all these conditions were considered. An example relative risk model for age-related macular degeneration disease, which is created in accordance with the common format, is seen in Table 4. Although this common format is a powerful format, alternatively, a simpler format (e.g. the ids of the variants, the alleles of the variants and the allele's risk of developing the disease) can be produced and integrated into the program. However, in this case, there will be no need for the relative risk table, the third component of our program. The reason of this is that the relative risk table and the classification table will have the same number of elements. Therefore, the relative risk table should be disabled if the simple format mentioned is used.

TABLE IV. RELATIVE RISK MODEL FOR AMD

Number	**Genotypes (Four Fields)**				**Effect Size (Odds Ratio)**
	rs1061170	*rs1410996*	*rs10490924*	*rs9332739 or rs641153*	
	Type = Genotype	*Type = Number of risk alleles present*	*Type = Genotype*	*Type = Rare allele present or not*	
	Alleles = T,C	*Alleles = A,G*	*Alleles = G,T*	*Alleles = G,C&G,A*	
0	C|C	2	G|G	Yes	16.2
1	C|C	2	G|G	No	30.0
2	C|C	2	G|T	Yes	49.8
3	C|C	2	G|T	No	92.5
⋮	⋮	⋮	⋮	⋮	⋮

The upper side of the thick line in Table 4 is called header part, whereas the underside is called values part. In the header part, the ids of the variants or variants group, which variants are handled together, the type of the variant or variant group in the risk model ("Genotype", "Number of risk alleles present", "Rare allele present or not") and the alleles of the variants must be specified. In the values part, genotype combinations and each genotype combination's risk of developing the disease (odds ratio) are contained. Any relative risk model constructed according to the common format can be stored as excel spreadsheet or can be stored in database.

C. The Dynamic Program and Its Components

The features such as computing the disease risks of a large number of people simultaneously and classifying people according to their genetic characteristics, are very important for personalized medicine and preventive medicine. Therefore, considering these features, a dynamic program was developed which is compatible with the relative risk models (formed in accordance with the common format).

The developed program has basically two inputs. The first of them, as we have mentioned above, is the relative risk model of any disease, formed in accordance with the common format.

The second input is individuals' genotypes, regarding the variations associated with the disease. These personal genotypes, which are the second input of the program, are brought dynamically from the database. On the other hand, in systems / clinics where personal genotypes are organized as excel spreadsheets, the second input of the program may also be an excel spreadsheet. An example table for individuals and their genotypes, regarding age-related macular degeneration disease, is seen in Table 5. By taking these two inputs, our program classifies individuals based on their genotypes and computes the disease risks of them.

Our program is based on the basic classification approach, which is described in the upper section, and consists of four main components. These components are "Variant Array", "Classification Table", "Relative Risk Table" (different from the relative risk model), and "Hash-Table". Classification table and relative risk table are shown in Fig. 1 and in Table 7, respectively.

TABLE V. AN EXAMPLE TABLE OF INDIVIDUALS AND THEIR GENOTYPES

Individuals	**Genotype Values**				
	rs1061170	*rs1410996*	*rs10490924*	*rs9332739*	*rs641153*
Individual4	T|C	G|G	G|T	G|G	G|A
Individual9	C|T	G|G	G|T	G|G	G|G
Individual47	T|T	A|A	G|G	G|G	G|G
⋮	⋮	⋮	⋮	⋮	⋮

1) Variant Array: The first and the simplest one from the components constituting our program is "Variant Array". A sample variant array is seen in Table 6. Variant array, as is evident from its name, stores the information concerning the variations specified in the relative risk model. Also, each element of the variant array consists of two attributes. The first attribute holds the id of the variation. On the other hand, the second attribute is "Alleles" array and, this attribute is used in some formulas in the basic classification approach.

TABLE VI. A SAMPLE VARIANT ARRAY

Indices	**Values**
0	rs1061170, (T, C)
1	rs1410996, (A, G)
2	rs10490924, (G, T)
⋮	⋮

After the program reads the relative risk model, firstly, it stores the variations defined in the header portion of the relative risk model into the variant array. Both the order of the variations and the order of the alleles in the header portion are preserved during the recording process. If there are variations addressed together in the header portion, program stores them one by one into the variant array after decomposing. After the storage of whole variations present in the header portion into the variant array, variation-related operations are now performed only through the variant array and we do not have to read the relative risk model again and again.

2) Classification Table: The second component constituting our program is "Classification Table" and this table is basically used to store individuals (as classified based on their genotypes). A sample classification table is seen in Fig. 1. Classification table is actually an array and the indices of the array are the class numbers. The size of the array is found according to (2). The parameter "AV_N" in (2), namely, the number of variations associated with the disease, is the size of the variant array. Therefore, after all the variations specified in the header portion of the relative risk model are stored into the variant array, classification table is created dynamically during the execution of the program. Each element of the classification table comprises of an integer variable showing which group of the relative risk table corresponds to that class, a linked-list storing individuals who are in that class and who have the same genotype and, lastly, a string array storing the genotype combinations in open format. (Note that the third component, string array, is not shown in Fig. 1.)

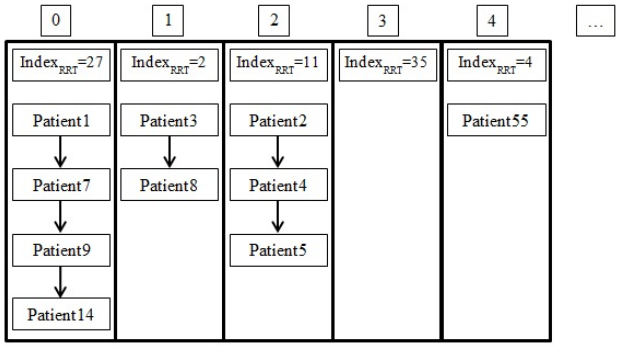


Fig. 1. A sample classification table

After the personal genotypes, which are the second input of the program, are fetched from the database, the Convert_to_Class_Number() algorithm, corresponding to (9), is executed for each individual and the classes of the individuals are determined. In addition, individuals are added to the list of classes, which they belong to. Other characteristics of the classification table are as follows: Array indices and integer variables (group number – index of relative risk table) establish a connection between classification table and relative risk table. On the other hand, by default, "-1" are assigned to the integer variables during the creation of the classification table. A value of -1 indicates that the class does not correspond to any group in the relative risk table. When relative risk table is being created, the values of the integer variables in the classification table are updated at the same time.

TABLE VII. A SAMPLE RELATIVE RISK TABLE SORTED BASED ON THE EFFECT SIZES

Indices (Group No)	**Values**		
	Effect Size (Odds Ratio)	*Genotype Preferences*	*List of Corresponding Classes*
0	1.0	(T|T, 0, G|G, Y)	1->2->3->4->5->6->7->8
1	1.9	(T|T, 0, G|G, N)	0
2	2.7	(T|T, 2, G|G, Y)	55->56->57->58->59->60->61->62
3	2.7	(T|T, 1, G|G, Y)	28->29->30->31->32->33->34->35
⋮	⋮	⋮	⋮

As mentioned before, in the case that a simpler format (e.g. the ids of the variants, the alleles of the variants and the allele's risk of developing the disease) is used for relative risk models, there will be no need for the relative risk table. Therefore, the relative risk table should be disabled and the elements (classes) of the classification table should be sorted based on the risk values if the simple format mentioned is used.

3) Relative Risk Table: The third component of our program is "Relative Risk Table". A sample relative risk table is seen in Table 7. Although relative risk table is very similar to the relative risk model, these two structures are distinct from each other. Relative risk models are arduous models that are generated by scientific research groups doing genome-wide association studies. These models show the disease-related variations and at what rate the genotypes influence the disease. On the other hand, relative risk table is a data structure that we developed, and it is a component of our program. Any relative risk model formed in compliance with the common format is converted into the corresponding relative risk table by our program. The relative risk table is also an array and the indices of the array are the group numbers. The size of this array is the same as the size of the relative risk model. Each element of the relative risk table consists of a float variable indicating the disease risk (Odds Ratio) of that group, an integer list storing which class/classes in the classification table that group corresponds to and a string array storing the combinations of the genotypes of that group.

4) Individual Hash Table: The fourth and the last component of our program is a hash-table, which stores individuals and classes they belong to. A sample hash-table is seen in Table 8. While explaining the classification table, we stated that the person's class is determined, and the person is added to the list of that class after the personal genotypes are fetched from the database. A similar process also applies to the hash table. Personal genotypes are brought from the database together with person information. In parallel, after the reading process, individuals, and classes they belong to are stored also in the hash table.

TABLE VIII. A SAMPLE HASH TABLE HOLDING INDIVIDUALS AND THEIR CLASSES

Indices	Values
0	(Individual1, 4)
1	
2	(Individual4, 13)
3	(Individual3, 2)
4	
5	(Individual9, 0)
⋮	⋮

We established a relation between the hash table and the classification table by storing individuals' classes in the hash table. We mentioned that there is a similar connection between classification table and relative risk table. Accordingly, both the connection between the hash table and the classification table and the connection between the classification table and the relative risk table are used to obtain the disease risk of any person. To find out the result of such a query, firstly, the person is searched in the hash-table and the class of the person is obtained, secondly, the relative risk table group corresponding to the person's class is determined from the classification table, and lastly, the disease risk of that group is obtained from the relative risk table. Since the person belongs to that group in the relative risk table, the risk of the person is the risk of that group. As depicted above, both the connections between the components and the hash-table significantly reduce the cost of search operations.

III. RESULTS AND DISCUSSION

The developed method and program were applied on the variation-based personal genetic data published by 1000 Genomes Project and the relative risk model generated for the age-related macular degeneration disease. The relative risk model that was used is seen in Table 4. After loading the data to the program, 2504 people were classified according to age-related macular degeneration disease. The classification table produced by the program is seen in Table 9 (Only a certain part of the classification table is shown here.).

TABLE IX. CLASSIFICATION TABLE AFTER AMD DISEASE RISK MODEL APPLIED ON THE SAMPLES OF 1000 GENOMES PROJECT

Class No	Genotype Preferences	Number of Individuals	Relative Risk No (Group No)	Risk
0	(T|T, A|A, G|G, G|G, G|G)	240	1	1.9
1	(T|T, A|A, G|G, G|G, G|A)	56	0	1
2	(T|T, A|A, G|G, G|G, A|A)	8	0	1
⋮		⋮	⋮	⋮
62	(T|T, G|G, G|G, C|C, A|A)	0	2	2.7
63	(T|T, G|G, G|T, G|G, G|G)	68	15	15.3
64	(T|T, G|G, G|T, G|G, G|A)	10	10	8.3
⋮		⋮	⋮	⋮
188	(C|C, A|A, T|T, C|C, A|A)	0	-1	-
189	(C|C, A|G, G|G, G|G, G|G)	1	-1	-
190	(C|C, A|G, G|G, G|G, G|A)	3	-1	-
⋮		⋮	⋮	⋮
240	(C|C, G|G, T|T, C|C, G|G)	0	33	154
241	(C|C, G|G, T|T, C|C, G|A)	0	33	154
242	(C|C, G|G, T|T, C|C, A|A)	0	33	154

In Table 9, the class number, the number of individuals in that class, the corresponding group number (in the relative risk table) and the risk of getting the disease for the individuals in that class are shown side by side. Note that the risk information is not normally included in the classification table, but it is shown here only for convenience. Depending on the number of variations in the relative risk model and (2), there are 243 separate classes in the classification table. Here, one remarkable case is that the group numbers corresponding to some classes is -1, that is, these classes do not correspond to any group (in the relative risk table). The emergence of this situation is due to the relative risk model (scientific research group's desire of simplifying the model as much as possible). For instance, although there are 243 different classes for this disease, there are only 36 groups in the relative risk model. Namely, the research group which created the model reduced

243 possible genotype combinations to 36 groups (depending on the allele frequencies in European population).

The relative risk table generated by the classification program is shown in Table 10. There are 37 groups in the relative risk table, and the first 36 groups correspond to 36 groups in the relative risk model created for AMD disease. The last group of the table contains classes that do not correspond to any group, and the number of this group is specified as -1. In addition, the list of classes corresponding to the relevant group, the number of individuals in that group, and the risk of that group are also indicated in Table 10.

TABLE X. RELATIVE RISK TABLE AFTER AMD DISEASE RISK MODEL APPLIED ON THE SAMPLES OF 1000 GENOMES PROJECT

Group No	**List of Classes**	**Genotype Preferences**	**Number of Total Individuals**	**Risk**
0	1->2->3->4->5->6->7->8	(T|T, 0, G|G, Y)	74	1
1	0	(T|T, 0, G|G, N)	240	1.9
2	55->56->57->58->59->60->61->62	(T|T, 2, G|G, Y)	19	2.7
⋮	⋮		⋮	⋮
33	235->236->237->238->239->240->241->242	(C|C, 2, T|T, Y)	4	154
34	153	(C|T, 2, T|T, N)	9	190
35	234	(C|C, 2, T|T, N)	6	285
-1	81->82->83->84->85->86-> . . .	-	20	-

The proposed program classifies individuals according to their genetic characteristics (the genotypes of the individual for the variations indicated in the relative risk model of the disease concerned) and produces the relevant outputs (tables). Finding the people who are genetically most similar to any person is a process that should be performed by the experts who use the program, not the program. This is because, for diagnostic or therapeutic purposes, different parameters and information may need to be considered in the process of finding the people who are genetically most similar to any person, and it is the specialist personnel who need to do this. On the other hand, even if there is no different parameter or information, experts should use the tables produced by the program and the relations between tables effectively in this process. From these perspectives, the decision maker is the expert staff, and the program only systematically presents the relevant information to the decision maker.

After the classification process is complete and the tables are generated, if the doctor wants to find the genetically most similar individuals to any individual, he or she should follow these steps: First, the person must be searched. When the person is found, the program will be positioned to the row in the hash table, where the person is located. At the same time, the program will be positioned to the row in the classification table, where the respective class is located. From the classification table, all persons in that class can be accessed, and besides, the genotype characteristics of that class, which group the class corresponds to in the relative risk table, and the relative risk value of the class can be obtained. The genotypes of all individuals in a class (regarding the variations indicated in the relative risk model) are exactly the same, that is, the individuals genetically most similar to each other are those in the same class. If there is no one else in the respective class, or if those present in the class do not provide adequate information about diagnosis or therapy, then the doctor will need to search in other classes. At this point, there are three options: 1) Other classes (if any) in the same group as the relevant class can be looked at in the relative risk table. 2) By making changes on the genotype characteristics of the relevant class, it can be easily found which class the new genotype combination corresponds to. 3) All other classes that have a different combination (1 allele different, 2 alleles different, and so on) from the genotype combination of the respective class can be easily found. The operations specified in the second and third options can be easily performed through the relevant parts of the program's interface. Thanks to the formulas and algorithms running in the background, the program returns the desired results simply and quickly. At this point, the only thing that the doctor must do is enter/select the values and press the button. As an example for the second option, assume that the doctor changed 2 alleles on the genotype combination (T|T, A|A, G|G, G|G, A|A) of class 2 and formed a new genotype combination as (C|C, A|A, G|G, G|G, A|A). If the doctor presses the corresponding button after entering the new genotype combination, the program will return that this new genotype combination corresponds to class 164. Besides, the program will be positioned to the row in the classification table, where class 164 is located. From the classification table, all individuals in class 164 can be accessed.

IV. CONCLUSION

In this study, a novel dynamic method and program is proposed for the disease-based genetic classification of individuals. Our generic classification method and program can classify individuals according to their disease-based genetic characteristics and can calculate disease risks of them simultaneously. The basic classification approach is completely based on the mathematical formulas and supports all types of variations. On the other hand, the common format was designed for the relative risk models, considering the common preferences of them. Our common format does not support relative risk models that include factors other than genetic factors, such as age, gender, smoking, etc. Our common format only supports relative risk models based on genetic factors. In parallel, the dynamic application, which is constructed on the basic classification approach, can work properly with the relative risk models developed for different diseases. Through this program, relative risk models can be managed from a single point, many people can be classified based on their genetic characteristics and the disease-risks can be calculated. On the other hand, calculating genetic similarities between classes is beyond the scope of this study. Depending on that, on the basis of disease, people who are genetically most similar to a person can be identified by only experts, using the outputs of the program (related tables). In short, this study contributes to personalized medicine approaches to some extent.

The relative risk model generated for the age-related macular degeneration disease and the personal variation data of 2504 people published by 1000 genomes project were applied to the program presented in this paper. With the loading of the relevant data to the program, 2504 people were classified according to age-related macular degeneration disease, the relative risks of these individuals were calculated, and the relevant tables, which are the outputs of the program, were produced. With the same logic, by using the relative risk

model for any disease, which is constructed in accordance with the common format, a large number of individuals can be classified according to the respective disease.

REFERENCES

[1] J. Hardy, A. Singleton, "Genomewide association studies and human disease", *New England Journal of Medicine*, *360*(17), 1759–1768, 2009.

[2] J. Krier, R. Barfield, R.C. Green, P. Kraft, "Reclassification of genetic-based risk predictions as GWAS data accumulate", *Genome medicine*, *8*(1), 1-11, 2016.

[3] Internet: GWAS Catalog, https://www.ebi.ac.uk/gwas/, 30.07.2021.

[4] T. A. Manolio, F. S. Collins, N. J. Cox, et. al., "Finding the missing heritability of complex diseases", *Nature*, *461*, 747–753, 2009.

[5] T. Beck, T. Rowlands, T. Shorter, A. J. Brookes, GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D986–D993, https://doi.org/10.1093/nar/gkac1017.

[6] Hettiarachchi, G., & Komar, A. A. (2022). Genome Wide Association Studies (GWAS) to Identify SNPs Associated with Common Diseases and Individual Risk. In Single Nucleotide Polymorphisms: Human Variation and a Coming Revolution in Biology and Medicine (pp. 51-76). Cham: Springer International Publishing.

[7] L. A. Hindorff, P. Sethupathy, H. A. Junkins, et. al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits", *Proceedings of the National Academy of Sciences*, *106*(23), 9362-9367, 2009.

[8] S. J. Schrodi, S. Mukherjee, Y. Shan, et. al., "Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future", *Frontiers in genetics*, *5*, 162, 2014.

[9] Lee, M. J., Lee, I., & Wang, K. (2022). Recent advances in RNA therapy and its carriers to treat the single-gene neurological disorders. Biomedicines, 10(1), 158.

[10] M. M. Alves, Y. Sribudiani, R. W. W. Brouwer, et. al., "Contribution of rare and common variants determine complex diseases-Hirschsprung disease as a model", *Developmental biology*, *382*(1), 320-329, 2013.

[11] J. Altmüller, L. J. Palmer, G. Fischer, et. al., "Genomewide scans of complex human diseases: True linkage is hard to find", *The American Journal of Human Genetics*, *69*(5), 936-950, 2001.

[12] J. C. Barrett, S. Hansoul, D. L. Nicolae, et. al., "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease", *Nature genetics*, *40*(8), 955-962, 2008.

[13] J. Maller, S. George, S. Purcell, et. al., "Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration", *Nature genetics*, *38*(9), 1055-1059, 2006.

[14] E. Zeggini, L. J. Scott, R. Saxena, et. al., "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes", *Nature genetics*, *40*(5), 638-645, 2008.

[15] K. Yasuda, K. Miyake, Y. Horikawa, et. al., "Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus", *Nature genetics*, *40*(9), 1092-1097, 2008.

[16] S. Kathiresan, B. F. Voight, S. Purcell, et. al., "Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants", *Nature genetics*, *41*(3), 334, 2009.

[17] Weeks, Elle M., et al. "Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases." Nature Genetics 55.8 (2023): 1267-1276.

[18] Abdellaoui, A., Dolan, C. V., Verweij, K. J., & Nivard, M. G. (2022). Gene–environment correlations across geographic regions affect genome-wide association studies. Nature genetics, 54(9), 1345-1354.

[19] C. Sabatti, S. K. Service, A. L. Hartikainen, et. al., "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population", *Nature genetics*, *41*(1), 35-46, 2009.

[20] W. Zheng, J. Long, Y. T. Gao, et. al., "Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1", *Nature genetics*, *41*(3), 324-328, 2009.

[21] C. Katsios, D. H. Roukos, "Individual genomes and personalized medicine: Life diversity and complexity", *Personalized Medicine*, *7*(4), 347-350, 2010.

[22] M. A. Hamburg, F. S. Collins, "The path to personalized medicine", *New England Journal of Medicine*, *363*(4), 301-304, 2010.

[23] G. S. Ginsburg, J. J. McCarthy, "Personalized medicine: Revolutionizing drug discovery and patient care", *TRENDS in Biotechnology*, *19*(12), 491-496, 2001.

[24] N. J. Schork, "Personalized medicine: Time for one-person trials", *Nature News*, *520*(7549), 609, 2015.

[25] Yamamoto, Y., Kanayama, N., Nakayama, Y., & Matsushima, N. (2022). Current status, issues and future prospects of personalized medicine for each disease. Journal of Personalized Medicine, 12(3), 444.

[26] Hassan, M., et. al. (2022). Innovations in genomics and big data analytics for personalized medicine and health care: A review. International journal of molecular Sciences, 23(9), 4645.

[27] The International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome", *Nature*, *431*(7011), 931-945, 2004.

[28] S. Levy, G. Sutton, P. C. Ng, et. al., "The diploid genome sequence of an individual human", *PLoS biology*, *5*(10), 2113–2144, 2007.

[29] International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs", *Nature*, *449*(7164), 851, 2007.

[30] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing", *Nature*, *467*(7319), 1061–1073, 2010.

[31] 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes", *Nature*, *491*(7422), 56-65, 2012.

[32] 1000 Genomes Project Consortium, "A global reference for human genetic variation", *Nature*, *526*(7571), 68-74, 2015.

[33] 1000 Genomes Project Consortium, "An integrated map of structural variation in 2,504 human genomes", *Nature*, *526*(7571), 75-81, 2015.

[34] Internet: 1000 Genomes Project Consortium, /vol1/ftp/release/20130502/ directory, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, 30.07.2021.

[35] Internet: 1000 Genomes Project Consortium, /vol1/ftp/release/20130502/supporting/bcf_files directory, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/bcf_files, 30.07.2021.

Journal of Emerging Computer Technologies
Research Article
Received: *2023-11-14* | ***Reviewing:*** *2023-11-15* & *2024-02-18* | ***Accepted:*** *2024-02-18* | ***Online:*** *2024-02-18* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1390428

Noise Removal from the Image Using Convolutional Neural Networks-Based Denoising Auto Encoder

Younus FAROOQ
Cankiri Karatekin University
Çankırı, Türkiye
younisfarooq321@gmail.com
0009-0006-4512-3989

Serkan SAVAŞ
Kırıkkale University,
Kırıkkale, Türkiye
serkansavas@kku.edu.tr
0000-0003-3440-6271

***Abstract*—With the exponential growth in the volume of digital images captured daily, there is an escalating demand for elevating image quality to achieve both accuracy and visual appeal. Addressing this need, the development of techniques for reducing image noise while preserving crucial features, such as edges, corners, and sharp structures, has become imperative. This paper delves into the significance of image denoising and introduces a novel approach utilizing a denoising autoencoder based on convolutional neural networks (CNNs). The proposed method adopts a meticulous two-step process to effectively eliminate noise. Initially, input images are segregated into training and testing sets. Subsequently, a denoising autoencoder model is trained using the designated training data. This model is then further refined through training on a CNN, enhancing its noise reduction capabilities. The evaluation of the system's performance is conducted using testing data to gauge its effectiveness. The study employs the MATLAB programming language for implementation and evaluation. Results, measured through RMSE (Root Mean Square Error) and PSNR (Peak Signal-to-Noise Ratio) criteria on two distinct datasets—the Covid19-radiography-database and SIIM-medical-images—reveal that our proposed method outperforms existing approaches significantly. This approach is particularly promising for applications demanding enhanced image quality, such as the resolution enhancement of medical images. The study contributes to the ongoing efforts in noise reduction research, offering a robust solution for improving visual perception in diverse image processing applications.**

Keywords—image noise, denoising autoencoder, convolutional neural network, image denoising.

I. INTRODUCTION

The surge in daily digital image capture has created a rising demand for images that are not only more accurate but also visually appealing. However, this surge in image capture also brings inevitable noise, diminishing overall image quality. Noise significantly impacts image quality in various applications like machine vision and object detection [1], as it can lead to false detections and inaccurate segmentations. Presently, most methods for noise removal primarily target grayscale image noise and struggle to effectively identify all compromised pixels. Various factors contribute to the noise that plagues digital images, including data transmission over noisy channels, hardware storage errors, and defective pixels during image capture [2]. Image denoising's primary objective is to preserve image structures, such as features, edges, and textures. Removing all types of noise before image analysis is crucial to prevent misinterpretation [3], [4]. An image denoising method's effectiveness hinges on how much noise it eliminates and how closely it preserves the original pixel values. Ineffective denoising can result in the loss of vital details, such as edge information. Over the past decades, experts have strived to develop efficient and accurate denoising techniques that reduce noise while maintaining essential visual characteristics [5]. Historically, image denoising methods relied on specific filters designed for certain distributions, rendering them less efficient when distribution characteristics weren't met. Recently, machine learning approaches have gained traction in noise reduction, with neural networks standing out. These algorithms attempt to predict the transformation of input data to output by learning from input-output pairs. Deep learning (DL) algorithms, which aim to mimic human observation, analysis, learning, and decision-making, have seen notable success in complex tasks, especially in diagnostics. DL's popularity stems from advances in on-chip processing, affordable hardware, and research in machine learning and signal processing [5]. DL techniques have particularly gained attention in image noise removal. Autoencoders, a type of neural network, aim to learn an approximation of the identity function using backpropagation [6]. Image denoising is another application of autoencoders, where they serve as non-linear functions to eliminate image noise. Fig. 1 illustrates an overview of an autoencoder.

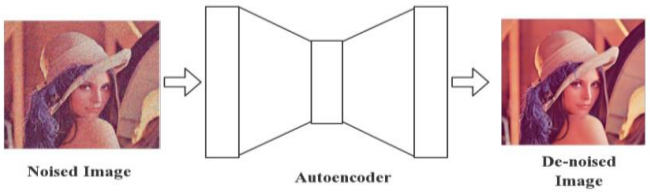


Fig. 1. A view of a denoising autoencoder [7]

In this research, a novel approach employs Denoising Autoencoders (DAE) in conjunction with Convolutional Neural Networks (CNN) to tackle image noise. Autoencoders have become a valuable framework for unsupervised learning of internal representations. This study utilizes a DL-optimized DAE model named CDAE for noise reduction. The CDAE model combines both CNN and DAE, offering a practical solution that works effectively regardless of the noise distribution in images. This network is trained by introducing random noise (specifically Gaussian noise) to the input image, with the objective of producing a noise-free original image as the output. This training strategy encourages the autoencoder to learn a function that removes noise and reconstructs the image. While existing image denoising methods have performed well, they suffer from drawbacks like manual

Cite (APA): Farooq, Y., Savaş, S. (2023). Noise Removal from the Image Using Convolutional Neural Networks-Based Denoising Auto Encoder. *Journal of Emerging Computer Technologies*, 3(1), 21-28. Doi: 10.57020/ject.1390428

parameter adjustments and specific model requirements. Recent advancements in DL, particularly the flexible CDAE architecture, have addressed these issues, making them more practical for real noisy images.

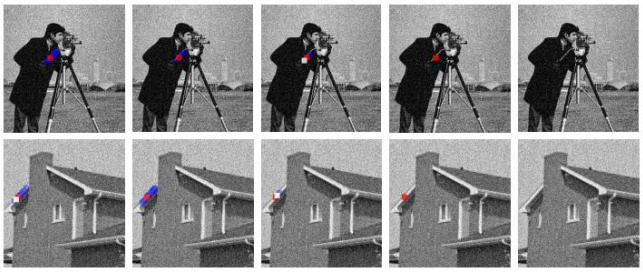


Fig. 2. The matching process of BM3D noise removal

The removal of noise from low-quality images is crucial in many fields, especially in the context of information sharing through digital images. Noise is an inevitable factor in image recognition, significantly degrading visual quality. Hardware issues, software operations like format conversion, copying, scanning, printing, and compression can introduce noise to images. This noise represents an unwanted signal recorded in the image. Consequently, noise removal is a fundamental step in image processing for various computer systems [8], [9]. Image denoising techniques aim to eliminate noise and restore image clarity. Distinguishing between noise, edges, and texture, all of which exhibit high-frequency components, poses a significant challenge in this field. Notably, the types of noise discussed in the literature encompass additive white Gaussian noise (AWGN), impulse noise, quantization noise, Poisson noise, and speckle noise [10]. AWGN originates from analogy circuits, while other noise types result from manufacturing flaws, bit errors, and low photon counts [11]. Image noise removal methods find application in diverse fields such as medical imaging, remote sensing, military surveillance, biometrics, forensics, industrial automation, agriculture, and human identification. In medical and biomedical imaging, denoising algorithms are crucial for eliminating noise types like speckle, Rician, and quantum noise. Remote sensing leverages noise reduction algorithms to address issues like AWGN and salt-and-pepper noise [12]. Military surveillance relies on synthetic aperture radar imagery, where denoising techniques have successfully reduced speckle artifacts. Image denoising techniques originated in the 1960s, initially employing two methods: transformation coefficients such as Fourier transform, discrete cosine transforms, and certain wavelets, as well as pixel value averaging. However, these methods often yielded substantial errors and adverse effects like excessive smoothing, stair-stepping, and ringing, resulting in reduced image quality [12]. In 2005, a novel approach called non-local averaging emerged. Unlike local averaging, which softens images by calculating pixel averages in the vicinity, this method computes the average of all image pixels using patches. Weights are assigned based on similarity to the desired pixel. This approach, exemplified by the block-matching and 3D filtering (BM3D) method, enhances image clarity and retains more details compared to local averaging [13]. Fig. 2 illustrates the process, where a 2D image is treated as 3D. It begins by estimating a noise-reduced image and then iteratively refines noise removal by assessing pixel similarities. Consequently, the image quality heavily relies on pixel similarity, yielding better results when pixels exhibit greater similarity [13].

In this context, the structure of the study is as follows. The second section describes the related studies, and the third section explains the methodology and materials applied. The fourth section describes the results obtained and the fifth section discusses these results. The sixth section concludes the study.

II. Related Works

Efficient learning models were introduced, capable of directly deriving the desired output from input data while conserving energy. Initially, these DL networks processed images in small patches, gradually enhancing results through architectural improvements, advanced cost functions, and newer activation functions. Kamal Bajaj and colleagues introduced a deep learning model based on autoencoders for image denoising, focusing on learning noise from training images to produce clean images [14]. The architecture comprises layers of Convolution, Pooling, Deconvolution, and up sampling, forming a self-encrypting block with a total of 15 layers. Two key objectives guide the establishment of connections between layers: increasing depth to extract more image features and preventing gradient disappearance during network training for improved image reconstruction. Performance evaluation metrics include signal-to-noise ratio and structural similarity index. [15] developed a large-scale denoising convolutional neural network (DnNCC) to remove JPEG compression noise, demonstrating high artifact removal performance with an improved learning algorithm. This technique aids in reducing artifacts, including motion artifacts in diagnostic images. [16] introduced the denoising autoencoder (DAE), demonstrating superior performance compared to traditional examples. [17] devised an adaptive multi-column deep neural network (DNN) with multi-stack sparse DAEs (SSDAE) to handle images corrupted by three types of noise. [18] combined denoising autoencoders (DAEs) and convolutional autoencoders (CAEs) for medical image denoising. [19] introduced a cumulative denoising autoencoder (SCDAE) with a hierarchical structure, embedding whitening layers to process input feature maps. [20] utilized a convolutional denoising autoencoder (CDAE) followed by a DAE in a cascaded fashion to address images with massive noise, highlighting the need for robust image classification systems that perform well across variable noise levels without extensive training.

In the Noise2Noise algorithm [21], the network is trained to perform image denoising solely based on noisy data, without any knowledge of the ground truth. This concept is further extended by the Noise2Void algorithm [22], which eliminates the necessity for pairs of noisy images during training. This feature is particularly pertinent in biomedical applications where ground truth images may be unavailable. A self-supervised approach is introduced by the Noise2Self method [23], obviating the need for prior information on the input image, noise estimation, or ground truth data. Image denoising [24] is accomplished through a convolutional neural network (CNN) extracting features from the noisy image. This method incorporates both edge regularization and total variation regularization. The combination of CNN and low-

rank representation is deployed to identify anomalous pixels in hyperspectral images [25]. For restoring blurred images affected by Cauchy noise, a multilevel wavelet convolutional neural network is applied [26]. The Block Matching Convolutional Neural Network (BM-CNN) [27] integrates deep learning with the 3D block-matching method, predicting the denoising of stacks through a DnCNN [28] trained on a dataset comprising 400 images, corresponding to over 250,000 training samples. A feed-forward Convolutional Neural Network is employed for smoothing images independently of the noise level, utilizing residual learning and batch normalization. Subsequently, the blocks are aggregated, and the image is reconstructed, akin to the 3D block-matching algorithm.

Researchers in [15] proposed the attention-directed CNN (ADNet) for image denoising, consisting of four blocks scattering (SB), feature enhancement (FEB), attention (AB), and reconstruction (RB) totaling 17 layers. SB, with 12 layers of Dilated Conv + BN + ReLU and Conv + BN + ReLU, enhances the effectiveness, performance, and depth reduction of the denoising framework. FEB incorporates three types of layers (Conv + BN + ReLU, Conv, and Tanh), and AB is a single convolution layer. [29] introduced the Noise Estimation Removal Network (NERNet) for noise reduction, consisting of modules for noise estimation and noise cancellation. The architecture integrates symmetric dilated blocks and pyramid feature fusion, adjusting to the noise level map for effective noise reduction. Gai and Bao in [30] utilized an upgraded CNN, MP-DCNN, for adaptive residual denoising. The model uses Leaky ReLU for noise extraction, employs SegNet for edge information retrieval, and utilizes MSE and perceptual loss for image reconstruction. Zhang et al. [31] proposed a dictionary learning model for mixtures of Gaussian distribution (MOG), adopting a minimization problem with sparse coding, dictionary updating, and hierarchical mapping functions to address the vanishing problem. Also, they proposed SANet which employed band aggregation, deep mapping, and convolutional separation blocks for noise removal. The architecture divides input noise into smaller blocks, maps and conceals each band, and aggregates all maps to create the output. Li et al. [32] suggested a detail-preserving CNN (DRCNN) that focuses on integrating high-frequency image material. DRCNN includes Generalization Module (GM) and Detail Preserving Module (DRM), lacking batch normalization, and addresses a detail loss function minimization problem. Xu et al. [33] introduced Bayesian deep matrix factorization (BDMF) for multiple image denoising, utilizing deep neural networks (DNN) for low-rank components and optimization through stochastic gradient variation Bayes. Jin et al. [34] proposed a classifier/regression CNN for image denoising, with a classifier network detecting impulse noise and a regression network restoring noisy pixels based on the classifier's prediction. Fang and Zeng [24] suggested the CNN variation model (CNN-VM) for picture denoising, employing EdgeNet with multiple scale residual blocks (MSRB) and edge regularization for feature extraction. Total variation regularization enhances shape edge performance, and Bregman splitting technique is used for solution discovery.

There are also innovative methods were also employed for network architecture, with residual learning being noteworthy as it focused on isolating noise rather than noise-free images. However, a fundamental issue persisted training these models required specificity to noise types and levels. Any change in the target noise necessitated a complete retraining from scratch. Deep neural networks have gained remarkable attention due to their exceptional performance in image-related tasks. Yet, the extended training duration, the challenge of hyperparameter selection, and other complexities inherent to DL cannot be overlooked. Ongoing efforts, such as batch normalization, aim to address these issues. Another drawback of DL methods is their limited performance when dealing with untrained noise types. For example, while BM3D can mitigate noise in images with mixed noise types, DL networks trained on specific noise types like Gaussian noise struggle with such scenarios. This limitation can be partly alleviated by training deep networks on more diverse real-world data [13].

III. Material and Methods

In this section, the different dataset groups used in the study are described, the proposed methodology is explained, and the evaluation metrics are presented.

A. Datasets

The proposed denoising system employs two medical datasets for training and testing, carefully selected to align with the objectives of this study. The first dataset utilized for evaluation is the chest X-ray (CXR) dataset [35], which comprises a variety of images representing different categories such as COVID-19, SARS, ARDS, and Streptococcus. Fig. 3 represent samples form the first dataset.

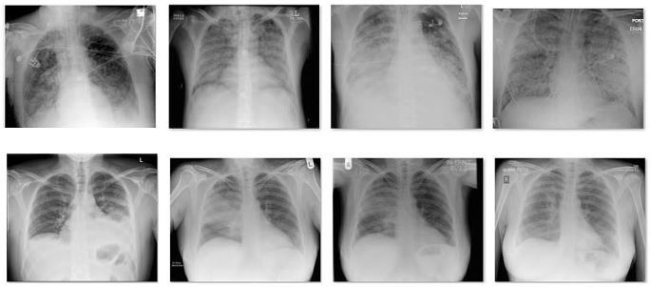


Fig. 3. Samples from the first dataset

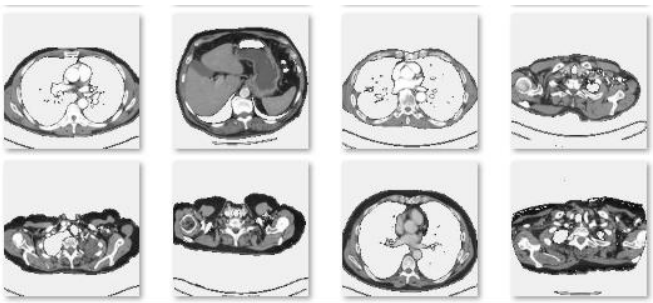


Fig. 4. Samples from the second dataset

In the second set of experiments, we utilize the CT Medical Images dataset [36], a subset of archived images from cancer imaging. These images are extracted from the central portion of computed tomography (CT) images and encompass a total of 475 images sourced from 69 distinct patients. This dataset was curated to assess various techniques for investigating trends in CT image data, specifically in relation to contrast

utilization and patient age. Fig. 4 represent samples from this dataset.

B. Proposed Method

The process involves four key stages. Initially, images are categorized into training and testing sets, where training data is used to train the model, and test data assesses its performance. The model undergoes training through an automatic coding method and the training dataset. Next, the data is fed into a CNN for additional learning. Finally, the testing data evaluates the system's performance. The system comprises four main blocks: the original image, noisy image generation, denoising via an autoencoder, and denoising using a CNN, each playing a vital role in the noise reduction process. we have detailed the steps of the proposed method in Fig. 5.

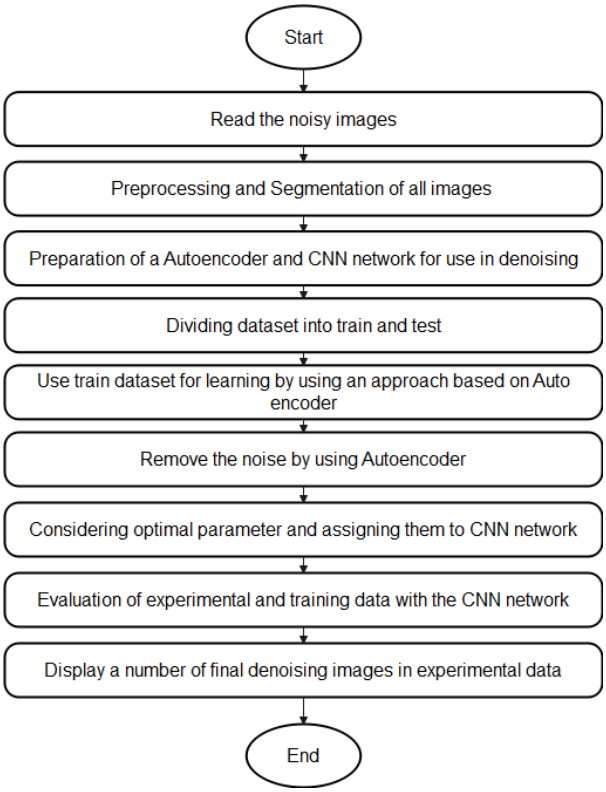


Fig. 5. The proposed method

This first step involves the selection of the database for processing and applying denoising using the CNN. In the second stage of this process, a crucial step unfolds as all the images undergo comprehensive preprocessing to ensure they are suitably prepared for subsequent analysis and processing. This meticulous preparation sets the foundation for effective data manipulation. Following this, in the third step, the groundwork is laid for the utilization of an autoencoder network, which plays a pivotal role in the denoising of images. This network, a cornerstone of the denoising process, is prepared with great care and precision. Autoencoder aims to reduce dimensionality and discover features in the data. Constraints on hidden units prevent the model from learning identity mappings, and it's trained to predict its input. The basic structure, illustrated in Fig. 6, includes input (x), encoding (y) via an encoder (f), and reconstruction (r).

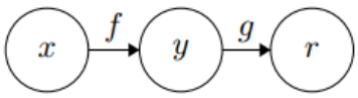


Fig. 6. General view of an auto-encryptor

The denoising autoencoder is designed to resist noise better by capturing high-level features and remaining robust to small input variations. It takes noisy input data and aims to produce clean, noise-free data. In our study, we introduce input noise by varying sample values before training and use CNNs. The process involves two steps: first, the model learns from the dataset using an autoencoder-based approach, and then noise is removed using another autoencoder. The output of this step is fed into a CNN to enhance image quality. The description and parameters of CNN layers and Auto Encoders are describing in Table I.

TABLE I. DESCRIPTION OF CNN LAYERS AND AUTO ENCODER PARAMERERS

Part	Parameter	Value
Images	Images size	120×120×3
CNN	layers CNN	imageInputLayer convolution2dLayer batchNormalization leakyReluLayer fullyConnectedLayer softmaxLayer classificationLayer
	Max Epochs	300
	ValidationFrequency	30
	Batch size	16
Auto Encoder	Number of Hidden Neurons	500
	Activation Function	Sig
	Ratio of noising features	0.4

According to Table I, the description outlines the parameters for a CNN and an Auto Encoder. The image-related parameters for the CNN include an image size of 120×120×3 pixels. The CNN architecture involves layers such as imageInputLayer, convolution2dLayer, batchNormalization, leakyReluLayer, fullyConnectedLayer, softmaxLayer, and classificationLayer. Training parameters for the CNN include a maximum of 300 epochs, validation frequency set to every 30 epochs, and a batch size of 16. The Auto Encoder specifications include 500 hidden neurons, a sigmoid activation function, and a 0.4 ratio for noising features.

Moving along to the fourth step, a pivotal division takes place within the dataset itself. The dataset is thoughtfully categorized into two subsets: the training dataset and the test dataset. This categorization is essential for systematically training and evaluating the performance of the proposed system. In the fifth step, an ingenious approach grounded in autoencoder principles is employed. This approach facilitates learning from the dataset, an integral part of the denoising process. The knowledge gleaned from this step forms the basis for subsequent noise removal. Subsequently, in the sixth step, the expertise acquired through the autoencoder-based approach is harnessed to effectively remove noise from the images. This marks a critical advancement in enhancing the visual quality of the images. The seventh step ushers in the utilization of the denoised output as input for a CNN, further advancing the image processing and denoising journey. Continuing to the eighth step, the parameters and weightings of the network's layers are meticulously fine-tuned and applied

to the CNN. This step marks the calibration of the CNN for optimal performance. Subsequent to this calibration, in the ninth step, the system is subjected to rigorous testing using data from the designated testing section. Here, the system's performance is subjected to a thorough evaluation, and its capability as a tester for the proposed system is assessed. In the tenth step, the evaluation process extends to encompass the experimental and simulation data, which are rigorously scrutinized using the optimized CNN network. Finally, in the last step, the culmination of this meticulous process is celebrated as a selection of the final denoised images is presented. These images represent the successful outcome of the entire experimental endeavor, reflecting the power, and efficacy of the applied denoising techniques.

C. Evaluation Metrics

One of the key benchmarks employed to assess the effectiveness of noise reduction techniques is the utilization of root mean square error (RMSE) deviation, as defined in (1). This criterion, as computed through the provided equations, quantifies the extent of disparity between the original image and the noise-reduced image. It's evident that a lower value for this parameter signifies a superior outcome.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_{ij} - y_{ij})^2}{N}} \quad (1)$$

where N represents the total number of image pixels, with x_{ij} denoting the pixel value in the original image at position i and j, and y_{ij} indicating the pixel value in the denoised image at the same position.

Signal-to-noise ratio (SNR) stands out as a vital evaluation metric in noise removal studies. This research leverages SNR to gauge both the distortion levels within the denoised image relative to the original and the overall image quality. Essentially, this metric quantifies changes in pixel intensity between the original and noise-reduced images. A higher SNR value signifies reduced distortion in comparison to the original image, indicating superior image quality. Peak signal-to-noise ratio (PSNR), as determined by (2), serves as a key component in this evaluation.

$$PSNR = 10 \times log_{10} \frac{255^2}{MSE} \quad (2)$$

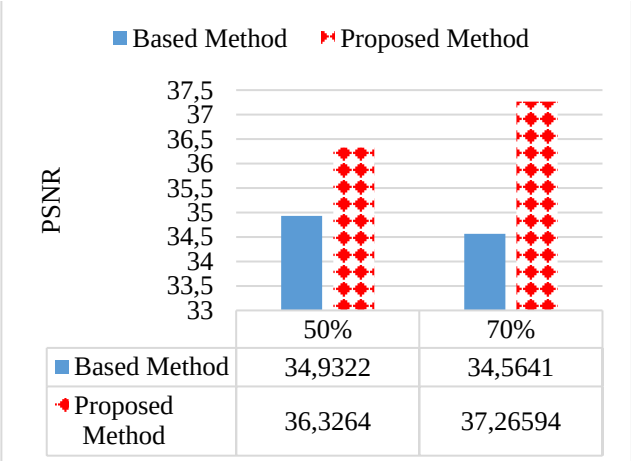
IV. RESULTS

The testing system boasts significant specifications. It runs on an 11th Gen Intel® Core™ i3-1115G4 processor, clocked at 3.00GHz. with 8.0 GB of RAM and a 260 GB SSD, it offers efficient memory and storage capabilities. Operating on Microsoft Windows 10 Ultimate, it provides a stable software environment. Additionally, MATLAB R2022a serves as the primary programming language, enabling versatile computational tasks. These specifications collectively establish the system's suitability for conducting research experiments and analyses.

We present the outcomes of our evaluations. To achieve this, we provide a comprehensive analysis of the evaluation results for each of the datasets introduced in the preceding section. Moreover, we will delve into the evaluation criteria, dissecting their impact on the results. It's important to note that we have considered the significance of training data volume for methods utilizing learning algorithms. As a result, we have incorporated the size of the training dataset as an evaluation parameter. Indeed, the results presented in this section encompass various training dataset sizes, specifically, 50% (with a corresponding testing size of 50%) and 70% (with a corresponding testing size of 30%). Our intent in varying the training dataset size is to discern the extent of influence on the compared methods as the training data volume fluctuates. Subsequently, we will proceed to delineate the evaluation results for each dataset, shedding light on the nuances uncovered during our analysis.

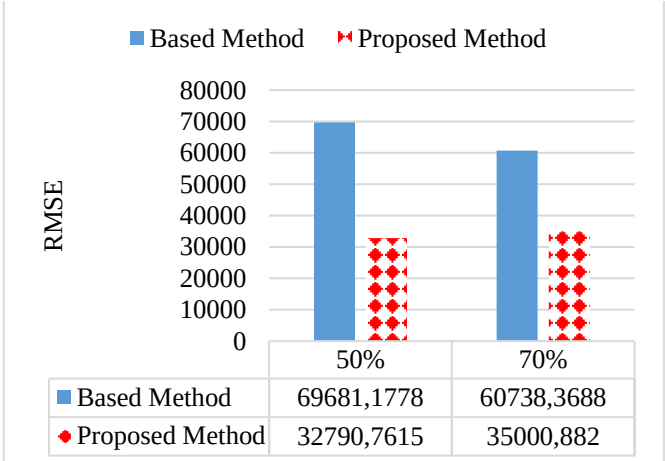
A. Results of the CXR Dataset

The initial set of findings within this section pertains to the assessment of PSNR and RMSE metrics on the CXR dataset. As a result, the results from the evaluation of both the proposed method and the base method based on these criteria are visually depicted in Fig. 7 and Fig. 8.



	50%	70%
Based Method	34,9322	34,5641
Proposed Method	36,3264	37,26594

Fig. 7. PSNR benchmark evaluation results on CXR dataset



	50%	70%
Based Method	69681,1778	60738,3688
Proposed Method	32790,7615	35000,882

Fig. 8. The evaluation results of the RMSE criterion on the CXR dataset

As depicted in Fig. 7 and Fig. 8, it becomes evident that the proposed approach outperforms the base method consistently. In essence, when evaluating the CXR dataset, the proposed method excels across both evaluation criteria and various training dataset sizes, yielding significantly superior results in comparison to the base method. A closer examination of the results laid out in this section unmistakably highlights the favorable performance of the proposed method. This is manifested in the notably lower error rates exhibited by the proposed method in contrast to the base method. Consequently, the images generated through the proposed method are of markedly superior quality. These findings underscore the robustness and effectiveness of the proposed approach, reinforcing its suitability for noise reduction tasks and its capacity to consistently deliver higher-quality results, particularly when confronted with varying training dataset sizes.

B. Results of CT Medical Images Dataset

In this section, similar to the preceding one, we will delve into the evaluation outcomes for each of the compared methods using the CT Medical Images dataset. These experiments encompass the inclusion of variations in the size of the training dataset as a dynamic parameter for the study. Within the ensuing Fig. 9 and Fig. 10, we present the evaluation results for each of the compared methods within this dataset.

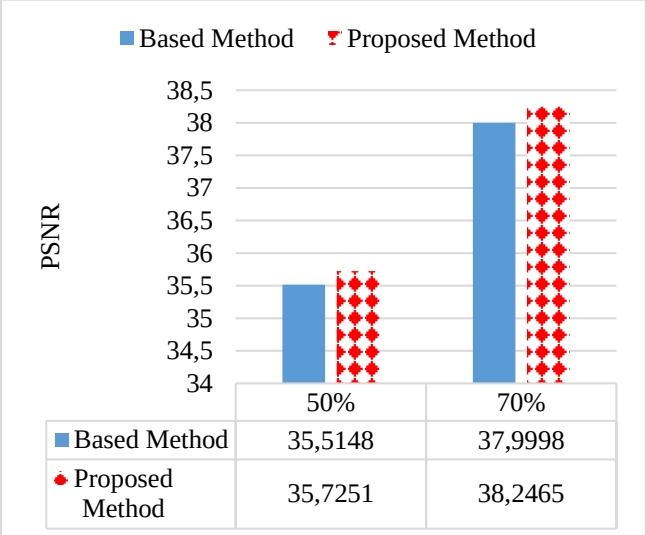


Fig. 9. PSNR benchmark evaluation results on CT medical images dataset

The results depicted in the aforementioned Fig. 9 and Fig. 10 substantiate that the proposed method outperforms the compared method when considering the PSNR criterion. Notably, variations in the size of the training dataset have failed to diminish the superiority exhibited by the proposed method. In fact, it is evident that the proposed method consistently outshines the compared method, even when the training dataset size is altered, as observed in the PSNR criterion.

However, a closer examination of the results presented in Fig. 10 reveals a divergence in performance between the two methods, with the compared method outperforming our approach in this specific criterion. This divergence is elucidated by the RMSE criterion, which indicates that the compared method yields lower error rates than the proposed method. A comprehensive summary of these evaluation results, underscoring the supremacy of the proposed method, is encapsulated in Table II. The ratios given in Table II are obtained by calculating the percentage of the difference between the proposed method and the base method to the base method.

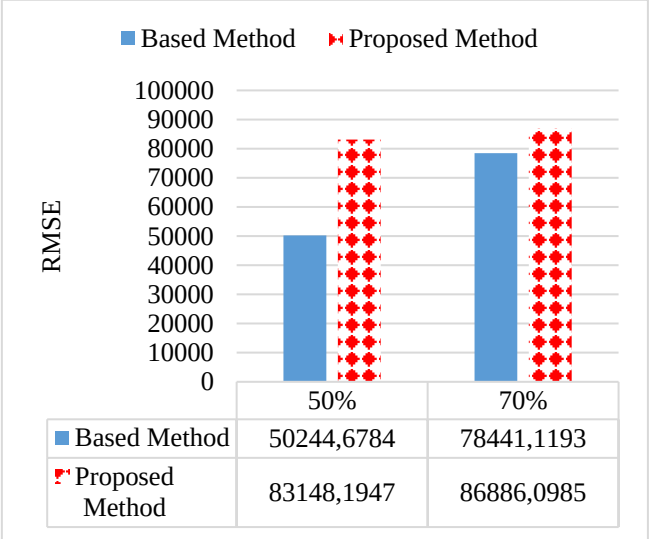


Fig. 10. RMSE evaluation results on the CT medical images dataset

TABLE II. THE IMPROVEMENT RATE OF THE PROPOSED METHOD COMPARED TO THE BASE METHOD

Dataset	Parameter	Comparison rates	
		50% - 50% Data Division	70% - 30% Data Division
CXR	PSNR	3.99%	7.82%
	RMSE	52.94%	42.37%
CT Medical Images	PSNR	0.59%	0.65%
	RMSE	-65.49%	-10.77%

Table II presents the comparison between the proposed method and the base method. In this comparison, a positive improvement, i.e. increase, is expected for PSNR and a negative improvement, i.e. decrease, is expected for RMSE. As can be seen in Table II, for the CXR data, both metrics are improved for both 50-50% data split and 70-30% data split. For CT medical images, although there is a slight improvement in the PSNR metric, there is no improvement in the RMSE metric, on the contrary, the proposed method gives worse results. According to these results, it can be stated that the auto encoder method fails for CT medical image but succeeds for CXR dataset.

In the case of the CT Medical Images dataset, the proposed method still achieves a commendable 5% improvement in PSNR, demonstrating its effectiveness in enhancing image quality. However, it's noteworthy that the RMSE criterion shows a negative improvement rate of -65%. While this may initially appear counterintuitive, it suggests that the proposed method, in certain scenarios, may yield slightly higher errors compared to state-of-the-art methods. Nonetheless, the combined results presented in Table 1 affirm the overall superiority of the proposed method in noise removal tasks across these datasets.

V. DISCUSSION

The results outlined in Table 1 not only highlight the significant effectiveness of the proposed noise removal method but also reveal its superiority over state-of-the-art approaches. The results substantial advancement underscores the method's prowess in achieving superior noise reduction quality. The consistent outperformance of the proposed method, especially when applied to the CXR dataset, signifies its robustness and reliability. Notably, this superiority remains resilient even when considering variations in the size of the training dataset, establishing it as a defining characteristic of the proposed method. This hallmark performance is indicative of the method's ability to adapt and maintain effectiveness across different data configurations.

The innovative dual-step approach to noise removal, integrating an autoencoder-based method for knowledge acquisition and subsequently employing autoencoder techniques in conjunction with a CNN, emerges as a pivotal factor in the success of the proposed method. This multi-step process plays a critical role in maximizing the quality of the resulting images, ensuring not only the removal of noise but also the preservation of essential image features. By strategically combining these techniques, the proposed method excels in noise reduction tasks, demonstrating a nuanced and sophisticated approach to addressing the challenges associated with medical image denoising.

It is crucial to emphasize the broader implications of the proposed method in the context of medical imaging. The superior performance observed on the CXR dataset suggests that the method holds significant promise for applications such as respiratory and cardiac imaging, where image quality is paramount for accurate diagnosis and treatment planning. Moreover, the consistent outperformance across datasets underscores the method's versatility, making it a robust candidate for a wide range of medical imaging modalities.

The findings of this study contribute not only to the advancement of medical image denoising but also provide valuable insights for future research directions. The success of the dual-step approach opens avenues for exploring similar strategies in other image processing tasks, fostering innovation in the broader field of computer vision. Additionally, the discussion prompts consideration of potential refinements or extensions to the proposed method, such as exploring variations in the architecture or incorporating additional layers to further enhance its adaptability and generalization capabilities.

VI. CONCLUSION

DL studies have been successfully applied in many different fields, especially in the last decade. Image processing is one of the most common among these fields. One of the most important limitations for DL and image processing studies applied in many different fields such as health [37], [38], education [39], [40], communication [41], industry, agriculture [42] etc. is the noise in the data. Noise in medical images arises from various sources, including transmission and environmental factors, resulting in types like Gaussian, Poisson, blur, speckle, and salt-and-pepper noise. Noise reduction is vital in medical imaging, with filters like median, Gaussian, and Wiener customized for specific noise types. However, no universal solution meets all medical image denoising needs.

This study presents a tailored medical image noise reduction method, Automatic Noise Removal via Convolutional Neural Networks, using a two-step algorithm. It categorizes images into training and testing sets and employs automatic coding on the training data, training a CNN. Testing data evaluates the system. Efficacy was assessed using MATLAB, and a basic CNN method was implemented. Evaluation results, based on RMSE and PSNR criteria on datasets, clearly affirm the proposed method's consistent superiority over the base method.

This research highlights the effectiveness of CNN architectures in noise removal from images. The second chapter offers an overview of various CNN-based image denoising techniques, acknowledging both strengths and limitations. Challenges include limited memory capacity for CNN programs and the complexity of unsupervised denoising tasks. Furthermore, CNN methods remain relatively underutilized in medical image denoising. Future research may explore memory allocation for CNN tasks and the integration of these methods into expanding computer systems for disease diagnosis. Additionally, researchers in this field could draw inspiration from the approach presented in this study to devise novel image denoising methods, potentially by combining it with existing techniques for enhanced noise reduction.

REFERENCES

[1] A. A. Saraiva, M. S. de Oliveira, P. B. de Moura Oliveira, E. J. Solteiro Pires, N. M. Fonseca Ferreira, and A. Valente, "Genetic algorithm applied to remove noise in DICOM images," J. Inf. Optim. Sci., vol. 40, no. 7, pp. 1543–1558, 2019, doi: 10.1080/02522667.2019.1597999.

[2] S. Rani, Y. Chabrra, and K. Malik, "An Improved Denoising Algorithm for Removing Noise in Color Images," Eng. Technol. Appl. Sci. Res., vol. 12, no. 3, pp. 8738–8744, 2022, doi: 10.48084/etasr.4952.

[3] D. G. Kim, M. Hussain, M. Adnan, M. A. Farooq, Z. H. Shamsi, and A. Mushtaq, "Mixed Noise Removal Using Adaptive Median Based Non-Local Rank Minimization," IEEE Access, vol. 9, pp. 6438–6452, 2021, doi: 10.1109/ACCESS.2020.3048181.

[4] A. Mukherjee, S. Sarkar, and S. K. Saha, "Segmentation of natural images based on super pixel and graph merging," IET Comput. Vis., vol. 15, no. 1, pp. 1–11, 2021, doi: 10.1049/cvi2.12008.

[5] A. Jindal, N. Aggarwal, and S. Gupta, "An Obstacle Detection Method for Visually Impaired Persons by Ground Plane Removal Using Speeded-Up Robust Features and Gray Level Co-Occurrence Matrix," Pattern Recognit. Image Anal., vol. 28, no. 2, pp. 288–300, 2018, doi: 10.1134/S1054661818020086.

[6] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun., pp. 278–282, 2018, doi: 10.1109/ICSCCC.2018.8703316.

[7] Y. Zhang, "A Better Autoencoder for Image: Convolutional Autoencoder," pp. 1–7, 2015.

[8] A. Semwal, A. Chamoli, and A. Semwal, "A SURVEY : On Image Denoising And Its Various Techniques," Int. Res. J. Eng. Technol., pp. 1565–1568, 2017.

[9] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-Based image denoising," IEEE Trans. Image Process., vol. 27, no. 9, pp. 4608–4622, 2018, doi: 10.1109/TIP.2018.2839891.

[10] A. E. Ilesanmi and T. O. Ilesanmi, "Methods for image denoising using convolutional neural network: a review," Complex Intell. Syst., vol. 7, no. 5, pp. 2179–2198, 2021, doi: 10.1007/s40747-021-00428-4.

[11] B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, "Image denoising review: From classical to state-of-the-art approaches," Inf. Fusion, vol. 55, pp. 220–244, 2020, doi: 10.1016/j.inffus.2019.09.003.

[12] S. Sudha, G. R. Suresh, and R. Sukanesh, "Speckle Noise Reduction in Satellite Images Using Spatially Adaptive Wavelet Thresholding," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 2, pp. 3432–3435, 2012.

[13] A. A. A. Goshtasby, "Advances in Computer Vision and Pattern Recognition," Image Regist. Princ. Tools methods, pp. 7–66, 2012, [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-4471-2458-0

[14] K. Bajaj, D. K. Singh, and M. A. Ansari, "Autoencoders Based Deep Learner for Image Denoising," Procedia Comput. Sci., vol. 171, pp. 1535–1541, 2020, doi: 10.1016/j.procs.2020.04.164.

[15] P. Svoboda, M. Hradis, D. Barina, and P. Zemcik, "Compression artifacts removal using convolutional neural networks," J. WSCG, vol. 24, no. 2, pp. 63–72, 2016.

[16] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," Proc. 25th Int. Conf. Mach. Learn., pp. 1096–1103, 2008, doi: 10.1145/1390156.1390294.

[17] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," Adv. Neural Inf. Process. Syst., 2013.

[18] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," IEEE Int. Conf. Data Min. Work. ICDMW, vol. 0, pp. 241–246, 2016, doi: 10.1109/ICDMW.2016.0041.

[19] B. Du et al., "Stacked Convolutional Denoising Auto-Encoders for Feature Representation," IEEE Trans. Cybern., vol. 47, no. 4, pp. 1017–1027, 2017.

[20] S. S. Roy, S. I. Hossain, M. A. H. Akhand, and K. Murase, "A robust system for noisy image classification combining denoising autoencoder and convolutional neural network," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 224–235, 2018, doi: 10.14569/IJACSA.2018.090131.

[21] J. Lehtinen et al., "Noise2Noise: Learning image restoration without clean data," 35th Int. Conf. Mach. Learn. ICML 2018, vol. 7, pp. 4620–4631, 2018.

[22] A. Krull, T. O. Buchholz, and F. Jug, "Noise2void-Learning denoising from single noisy images," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 2124–2132, 2019, doi: 10.1109/CVPR.2019.00223.

[23] J. Batson and L. Royer, "Noise2Seif: Blind denoising by self-supervision," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 826–835, 2019.

[24] Y. Fang and T. Zeng, "Learning deep edge prior for image denoising," Comput. Vis. Image Underst., vol. 200, 2020, doi: 10.1016/j.cviu.2020.103044.

[25] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, and Q. Li, "Hyperspectral Anomaly Detection via Deep Plug-and-Play Denoising CNN Regularization," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 11, pp. 9553–9568, 2021, doi: 10.1109/TGRS.2021.3049224.

[26] T. Wu, W. Li, S. Jia, Y. Dong, and T. Zeng, "Deep multi-level wavelet-CNN denoiser prior for restoring blurred image with cauchy noise," IEEE Signal Process. Lett., vol. 27, pp. 1635–1639, 2020, doi: 10.1109/LSP.2020.3023299.

[27] B. Ahn and N. I. Cho, "Block-Matching Convolutional Neural Network for Image Denoising," 2017, [Online]. Available: http://arxiv.org/abs/1704.00524

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3142–3155, 2017, doi: 10.1109/TIP.2017.2662206.

[29] B. Guo, K. Song, H. Dong, Y. Yan, Z. Tu, and L. Zhu, "NERNet: Noise estimation and removal network for image denoising," J. Vis. Commun. Image Represent., vol. 71, 2020, doi: 10.1016/j.jvcir.2020.102851.

[30] S. Gai and Z. Bao, "New image denoising algorithm via improved deep convolutional neural network with perceptive loss," Expert Syst. Appl., vol. 138, 2019, doi: 10.1016/j.eswa.2019.07.032.

[31] L. Zhang, Y. Li, P. Wang, W. Wei, S. Xu, and Y. Zhang, "A separation–aggregation network for image denoising," Appl. Soft Comput. J., vol. 83, 2019, doi: 10.1016/j.asoc.2019.105603.

[32] X. Li et al., "Detail retaining convolutional neural network for image denoising," J. Vis. Commun. Image Represent., vol. 71, 2020, doi: 10.1016/j.jvcir.2020.102774.

[33] D. Yin et al., "Speckle noise reduction in coherent imaging based on deep learning without clean data," Opt. Lasers Eng., vol. 133, 2020, doi: 10.1016/j.optlaseng.2020.106151.

[34] L. Jin, W. Zhang, G. Ma, and E. Song, "Learning deep CNNs for impulse noise removal in images," J. Vis. Commun. Image Represent., vol. 62, pp. 193–205, 2019, doi: 10.1016/j.jvcir.2019.05.005.

[35] "https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database."

[36] "https://www.kaggle.com/datasets/kmader/siim-medical-images".

[37] R. Butuner, & M. H. Calp. Diagnosis and Detection of COVID-19 from Lung Tomography Images Using Deep Learning and Machine Learning Methods. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 190-200, 2022.

[38] B. Guler Ayyildiz, R. Karakis, B. Terzioglu, & D. Ozdemir. Comparison of deep learning methods for the radiographic detection of patients with different periodontitis stages. Dentomaxillofacial Radiology, 53(1), 32-42, 2024.

[39] Q. Gazawy, S. Buyrukoglu and A. Akbas, "Deep Learning for Enhanced Education Quality: Assessing Student Engagement and Emotional States," 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), Sivas, Türkiye, 2023, pp. 1-8, doi: 10.1109/ASYU58738.2023.10296748.

[40] O. Güler, & İ. Yücedağ, Artırılmış Gerçeklik: Montaj ve Bakım Uygulamalarında El Tanıma Teknolojisi İle Etkileşim Çalışmaları. Paper presented at the 20. Akademik Bilişim Konferansı, Karabük. 2018. http://indexive.com/uploads/papers/pap_indexive15949797202147483647.pdf

[41] N. Daldal, Z. A. Sezer, M. Nour, A. Alhudhaif, K. Polat, A New Generation Communication System Based on Deep Learning Methods for the Process of Modulation and Demodulation from the Modulated Images, Mathematical Problems in Engineering, vol. 2022, Article ID 9555598, 13 pages, 2022. https://doi.org/10.1155/2022/9555598.

[42] R. Bütüner and M. H. Calp, (2023). Robotic Systems and Artificial Intelligence Applications in Agriculture. Current Studies in Technology, Innovation and Entrepreneurship, 145.

Journal of Emerging Computer Technologies
Research Article
Received: *2024-01-30* | ***Reviewing:*** *2024-02-01* & *2024-03-08*| ***Accepted:*** *2024-03-10* | ***Online:*** *2024-02-10* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1428146

Fraud Transaction Detection For Anti-Money Laundering Systems Based On Deep Learning

Jorge Felix Martínez Pazos
Xetid University of Informatics Science
Havana, Cuba
jorgefmp.mle@gmail.com
0009-0009-2477-8611

Jorge Gulín González
Center for Computational Mathematics Studies
University of Informatics Science
Havana, Cuba
jgulingonzalez@gmail.com
0000-0001-7912-2665

David Batard Lorenzo
Center for Computational Mathematics Studies
University of Informatics Science
Havana, Cuba
dbatardl@gmail.com
0009-0007-3555-2875

Jorge Alejandro Robaina Morales
Xetid University of Informatics Science
Havana, Cuba
jorgea.robaina.dev@gmail.com
0009-0000-5077-5718

Moises Miguel Rodríguez Álvarez
Department of Digital Systems
University of Informatics Science
Havana, Cuba
mrdguezalvz@gmail.com
0009-0001-4582-7875

***Abstract*— This study addresses the escalating problem of financial fraud, with a particular focus on credit card fraud, a phenomenon that has skyrocketed due to the increasing prevalence of online transactions. The research aims to strengthen anti-money laundering (AML) systems, thereby improving the detection and prevention of fraudulent transactions. For this study, a Dense Neural Network (DNN) has been developed to predict fraudulent transactions with efficiency and accuracy. The model is based on deep learning, and given the highly unbalanced nature of the dataset, balancing techniques were employed to mitigate the bias towards the minority class and improve performance. The DNN model demonstrated robust performance, generalizability, and reliability, achieving over 99% accuracy across training, validation, and test sets. This indicates the model's potential as a powerful tool in the ongoing fight against financial fraud. The results of this study could have significant implications for the financial sector, corporations, and governments, contributing to safer and more secure financial transactions.**

Keywords— AMLs, Deep Learning, Dense Neural Networks, Financial Fraud, Fraud Transaction Detection.

I. INTRODUCTION

The issue of financial fraud is becoming increasingly prevalent, with far-reaching implications for the finance sector, businesses, and governments. One particular area of concern is credit card fraud, which has seen a rise in incidence due to the growing popularity of online transactions. Credit card fraud can be classified into two categories: internal fraud, which involves collaboration between cardholders and banks using false identities, and external fraud, which involves the use of stolen credit cards. Traditional methods for detecting fraudulent transactions are often slow and ineffective. As such, financial institutions are now turning to computational approaches to tackle the problem of credit card fraud [1].

In the contemporary era, marked by significant scientific and technological advancements, financial and banking institutions have increasingly turned to Anti-Money Laundering Systems (AMLs). AMLs serve as a robust line of defense against illicit activities such as money laundering and terrorist financing, which pose substantial threats to the integrity of financial systems and the broader economy. The adoption of AMLs by financial institutions is driven by the need to comply with stringent regulatory requirements, protect customer data, and maintain the reputation of the institution. AMLs employ sophisticated algorithms and machine learning techniques to monitor transactions and user activities, identify suspicious patterns, and generate alerts for further investigation. The effectiveness of AMLs is enhanced by the integration of various components that work in synergy to provide a comprehensive and robust approach to detecting and preventing fraudulent transactions [2].

The first line of defense in AML is transaction monitoring, which involves the active tracking and analysis of financial transactions. This process, which can be performed in real-time or periodically, is designed to identify and prevent fraudulent or illegal activity. As technology advances, transaction monitoring is becoming increasingly automated and relies on machine learning.

In parallel with transaction monitoring, user activity monitoring logs and tracks user actions on devices, networks, or websites. This component of AML is critical for detecting and stopping insider threats, whether accidental or malicious.

To further enhance the effectiveness of AMLs, rule-based approaches are used, which capture the knowledge of a human expert in a specialized domain and embody it in a computer system. The rules, encoded in the system as if-then-else statements, provide a structured way to analyze transactions and user activity.

Another key component of AMLs is graph analytics, which involves the study and manipulation of data structures that encapsulate relationships between entities. The analysis facilitates the identification of patterns, anomalies, and structures within relational data that can be critical in detecting money laundering patterns such as smurfing, ring, cascade, in-out, and direct star. Graph theory algorithms such as centrality detection, community detection, and node similarity allow knowledge to be extracted from transaction and user behavior. These algorithms provide insight into the importance of nodes

Cite (APA): Martínez Pazos, J.F., Gulín González, J., Batard Lorenzo, D., Robaina Morales, J.A., Rodríguez Álvarez, M.M. (2023). Fraud Transaction Detection For Anti-Money Laundering Systems Based On Deep Learning. *Journal of Emerging Computer Technologies*, 3(1), 29-34. Doi: 10.57020/ject.1428146

within the network (centrality), the clustering of similar nodes (community detection), and the similarity between nodes (node similarity).

Finally, Know Your Customer (KYC) and Know Your Business (KYB) processes are regulatory requirements that ensure companies are doing business with legitimate individuals and entities. KYC focuses on identifying individual customers, while KYB verifies the companies they do business with. Both processes are critical to preventing online fraud and financial crime and complying with anti-money laundering regulations.

The adoption of AMLs by financial and banking institutions represents a proactive and strategic response to the challenges posed by financial fraud and cyber threats. By leveraging advanced technologies and scientific knowledge, these institutions are better equipped to detect and prevent fraudulent transactions, thereby contributing to the security and stability of financial systems.

Several approaches have been used to detect fraudulent transactions. Machine learning based approaches have been intelligently used to detect fraudulent transactions by analyzing a large number of financial data. The most satisfactory machine learning techniques such as an ensemble of decision tree (EDT), and deep learning techniques such as stacked auto-encoders (SAE) and restricted Boltzmann machines (RBM) classifiers are applied to the preprocessed data [3]. Varmedia et al, 2019 and Xuan et al, 2018 employed supervised learning approaches, mainly Random Forest, Logistic Regression, and Gaussian Naive Bayes, for the task of fraud detection in transactions, while Dornadula & Geetha, 2019 applied a sliding window technique to the data, which allowed them to aggregate multiple transactions over time, thereby obtaining a greater amount of important information about transactions over time. On the other hand, Jhon & Nazz, 2019 and Zadafiya et al. 2022 used unsupervised learning methods, specifically local outlier factors, and isolation forests, for fraud detection. These approaches are particularly practical in scenarios where it is necessary to assign a score to a transaction to determine its level of anomaly.

The existing financial infrastructure within Cuba is currently devoid of a proactive, automated Anti-Money Laundering (AML) system. This deficiency underscores the necessity for an in-depth exploration and understanding of the operational mechanisms of AML systems. Consequently, this research endeavor is designed to illuminate these mechanisms and propose an innovative solution specifically tailored to enhance the transaction monitoring facet of an AML system. This proposed solution aims to bolster the efficiency and effectiveness of detecting and preventing illicit financial activities, thereby fortifying the integrity of Cuba's banking system.

The following encapsulates the key contributions that underscore the significance of the forthcoming study:

- The development of a robust model, grounded in Neural Networks, that achieves transaction classification with an accuracy exceeding 99%.
- The implementation of the Min Class Balance reaffirms the potential of balanced training sets for enhancing the robustness and generalizability of Machine Learning and Deep Learning models. This approach effectively mitigates biases towards the less-represented class.
- A comprehensive initial guide covering most components of Anti-Money Laundering (AML) systems, addressing the current lack of information and documentation provided by companies and organizations regarding these solutions.

II. Materials & Methods

A. Dataset

The study in question utilizes a dataset, sourced from Kaggle, that is pivotal in the exploration of credit card fraud. This dataset is employed to construct an Artificial Neural Network (ANN) with the aim of efficiently and accurately predicting fraudulent transactions based on the dataset's features [9].

The `distance_from_home` feature, which measures the distance from the cardholder's residence to the transaction location, could potentially flag anomalous activity if transactions consistently occur at locations significantly distant from the cardholder's habitual transaction sites. The `distance_from_last_transaction` feature calculates the spatial difference between the current transaction and the preceding one. A substantial shift in transaction locations could potentially serve as an indicator of fraudulent activity. The `ratio_to_median_purchase_price` feature, representing the ratio of the transaction's purchase price to the median purchase price, could suggest fraudulent activity if a transaction deviates significantly from the median. The `repeat_retailer` feature indicates whether the transaction was conducted with a retailer previously used by the cardholder. Regular transactions with the same retailer could suggest a trusted retailer or a potential point of compromise. The `used_chip` feature signifies whether the transaction was executed using a chip-enabled credit card. Transactions conducted using a chip are generally deemed more secure than those using a magnetic stripe. The `used_pin_number` feature denotes whether the transaction was authenticated using a PIN. Transactions verified using a PIN are typically considered more secure. The `online_order` feature indicates whether the transaction was conducted online. Online transactions could potentially be more susceptible to fraud if adequate security measures are not implemented. Finally, the `fraud` feature signifies whether the transaction was fraudulent. This is likely the target variable for predictive modeling [9].

Each of these features contributes to the comprehensive understanding of credit card fraud. By studying patterns and anomalies in these features, it may be possible to construct a model that can accurately predict fraudulent transactions, thereby helping to mitigate the impact of credit card fraud. The insights derived from this analysis could be instrumental in enhancing the security measures employed by financial institutions and fostering a safer transaction environment for consumers. This study exemplifies the potential of machine learning in enhancing the security and reliability of financial systems and institutions.

B. Exploratory Data Analysis

The dataset under investigation, titled "Credit Card Fraud," comprises one million tuples and exhibits a significant imbalance, as illustrated in Figure 1. This imbalance is primarily due to the overwhelm-ing prevalence of genuine transactions compared to fraudulent ones. Such a disparity in class distribution poses unique challenges in model development, as the model must be sensitive enough to accurately identify the minority class (fraudulent transactions) without being overwhelmed by the majority class (genuine transactions). This aspect underscores the complexity of fraud detection in credit card transactions and highlights the need for sophisticated modeling techniques to effectively tackle this issue. The dataset initially contains 912,597 instances of genuine transactions and 87,403 instances of fraudulent transactions. However, to combat biases and enhance the model's performance and generalizability, a class balance is performed. This process adjusts the dataset so that genuine transactions constitute 60% of the total data and fraudulent transactions make up the remaining 40%. This approach ensures a better balance between the classes, which is crucial in machine learning models to prevent overfitting to the majority class and improve the detection of the minority class.

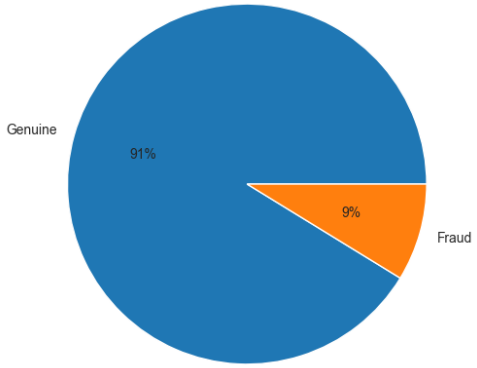


Figure 1. Pie chart of the dataset classes distribution

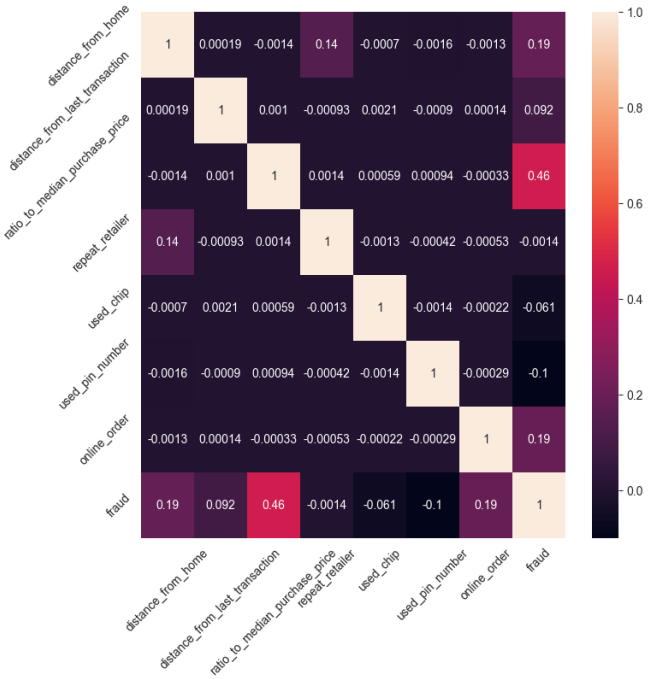


Figure 2: Correlational heatmap of the dataset attributes

As delineated in the correlational heatmap in figure 2, the feature most strongly correlated with whether a transaction is fraudulent is `ra-tio_to_median_price_purchase`. Despite this, other features such as `distance_from_home` and `online_order` exhibit a lesser degree of correlation. It's important to note that while these features may be less correlated, they still contribute valuable information that can enhance the predictive power of a model. In the complex landscape of fraud detection, even features with minor correlations can play a significant role when combined with other data points. Therefore, a comprehensive approach that considers a wide range of features is often most effective in accurately identifying fraudulent transactions.

C. Model Building & Training

Following the comprehensive data preprocessing and exploratory data analysis outlined above, the dataset has been partitioned into three distinct sub-sets: 70% for training, 15% for validation, and 15% for testing. This partitioning strategy facilitates the evaluation of the model's performance upon the completion of the training process.

The Dense Neural Network architecture designed for detecting fraud in transactions is constructed using TensorFlow's Keras API. This model is sequential, meaning that the layers are stacked linearly. The architecture begins with a dense layer with 64 neurons and a Rectified Linear Unit (ReLU) activation function. The input shape corresponds to the number of features in the training data (None, 7). The ReLU activation function is used to introduce non-linearity into the model, allowing it to learn more complex patterns. Following the first dense layer, a dropout layer is applied with a rate of 0.2. Dropout is a regularization technique that helps prevent overfitting by randomly setting a fraction of input units to 0 during training, which helps the model to generalize better to unseen data. The next layer is another dense layer with 32 neurons, again using a ReLU activation function. This is followed by another dropout layer with a rate of 0.2. The model then includes a third dense layer with 16 neurons and a ReLU activation function, followed by a dropout layer with a higher rate of 0.4. This increased dropout rate may help to further regularize the model and reduce overfitting. Next, a batch normalization layer is included. Batch normalization is a technique to provide any layer in a neural network with inputs that have zero mean/unit variance, which aids in overall network training [10], [11]. Finally, the architecture concludes with a dense output layer with 2 neurons, corresponding to the two classes (fraudulent and genuine transactions). The softmax activation function is used in this layer to output a probability distribution over the two classes, meaning the output can be interpreted as the model's confidence that the transaction is fraudulent or genuine.

The model is compiled with the Adam optimizer with a learning rate of 3e-4 and the sparse categorical cross-entropy loss function. The Adam optimizer is an adaptive learning rate optimization algorithm that's been designed specifically for training deep neural networks. The sparse categorical cross-entropy loss is suitable for multi-class classification problems. The metric used to evaluate the model during training is accuracy. The subsequent figure 3 provides a graphical

representation of the Dense Neural Network architecture, illustrating input flow and output flow.

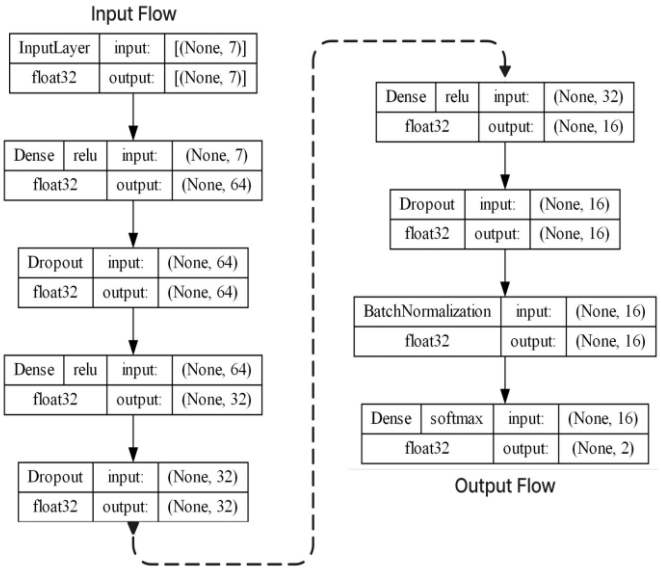


Figure 3. Architecture of the proposed Dense Neural Network model.

The model is trained using two specific callback functions: Early Stopping and Learning Rate Reduction. The Early Stopping callback is implemented to halt the training process when a monitored metric, in this case, validation accuracy, ceases to improve. The patience parameter is set to 10, indicating that the training will be stopped if there is no improvement in the validation accuracy after 10 epochs. The restore best weights parameter is set to True, ensuring that the model weights from the epoch with the optimal monitored metric are restored. The Learning Rate Reduction callback is utilized to reduce the learning rate when a metric has stopped improving. The metric monitored here is also the validation accuracy. The patience parameter is set to 4, denoting that if the validation accuracy does not improve after 4 epochs, the learning rate will be reduced. The factor parameter is set to 0.8, indicating that the learning rate will be reduced by a factor of 0.8. The min lr parameter is set to 0.000001, establishing the lower bound for the learning rate.

The model is then trained for a maximum of 150 epochs with a batch size of 1024, and using the validation set. The training process utilizes both the Early Stopping and Learning Rate Reduction callbacks [12]. This approach to training allows for a more efficient search for model parameters and can lead to improved model performance.

III. Results & Discussions

The model history of the implemented Dense Neural Network for fraud detection which is shown in Figure 4, offers an in-depth perspective on the model's learning progression and performance. The graphs depicting "Training & Validation Loss" and "Training & Validation Accuracy" collectively suggest that the model is effectively assimilating knowledge from the training data and demonstrating robust generalization capabilities when applied to unseen data in the validation set. This is substantiated by the consistent decrease in loss and increase in accuracy over time for both the training and validation sets, culminating in a training accuracy of 99.3%, validation accuracy of 99.7%, loss of 0.0202, and validation loss of 0.0075 at epoch 69. The near-perfect alignment of the loss and validation loss lines further indicates that the gradient is consistently moving toward an improved state.

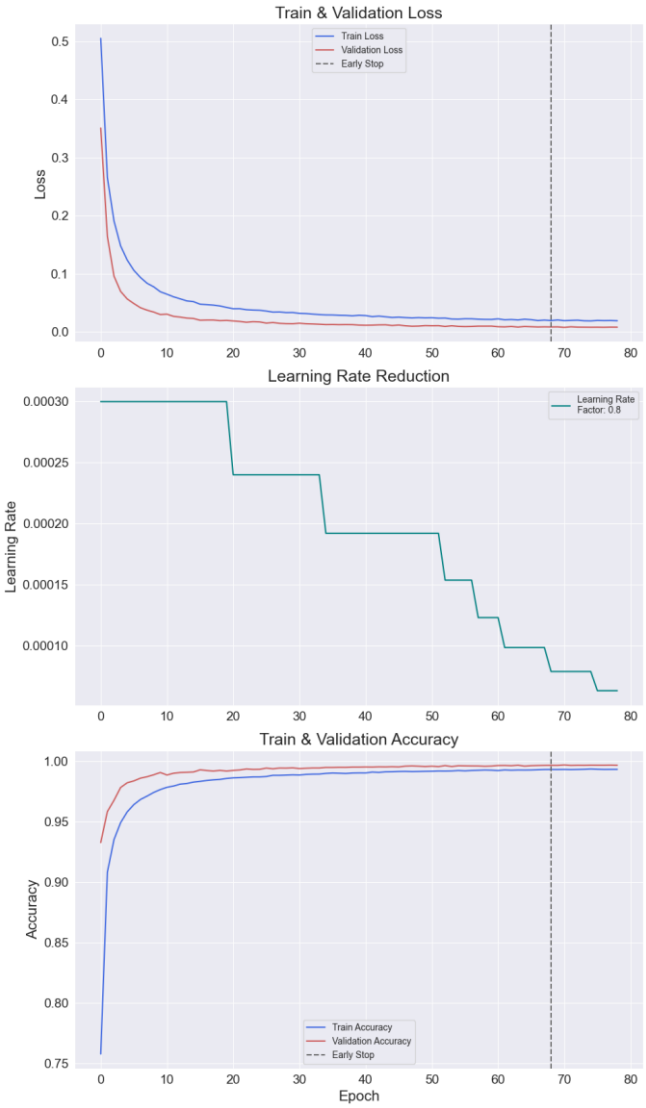


Figure 4. Model history of the proposed Dense Neural Network

The evaluation of predictive models on unseen data is of paramount importance to ascertain their per-formance in real-world scenarios. Without a com-prehensive evaluation on data that was not part of the training process, there exists a risk of sub-optimal real-world performance, potentially leading to erroneous decision-making. By subjecting the model to evaluation with novel data, a more accu-rate understanding of its real-world performance can be gleaned, thereby bolstering confidence in its deployment and ensuring its reliability and safety for integration into real-world solutions.

In the context of the Dense Neural Network model for fraud detection, the classification report provides a detailed account of its performance metrics [13]. The model exhibits a high degree of precision and recall for both ‘Fraud’ and ‘Not Fraud’ classes, with values exceeding 99% in all cases. Specifical-ly, the ‘Not Fraud’ class has a precision of 99.8% and a recall of 99.6%, resulting in an F1-score of 99.7%. Similarly, the ‘Fraud’ class has a precision of 99.4% and a

recall of 99.7%, leading to an F1-score of 99.6%. The overall accuracy of the model is 99.7%, with the macro and weighted averages for precision, recall, and F1-score all being 99.7%.

These results indicate that the model demonstrates a high level of effectiveness in distinguishing be-tween 'Fraud' and 'Not Fraud' instances. The high precision suggests that the model has a low false positive rate, while the high recall indicates a low false negative rate. The F1-score, being the harmonic mean of precision and recall, further confirms the model's robust performance. This comprehensive evaluation underscores the model's potential for reliable and safe deployment in real-world fraud detection solutions. Table 1 summarizes the classification report of the proposed Dense Neural Network.

Table 1: Summary of the classification report

	precision	recall	f1-score
Not Fraud	99.8	99.6	99.7
Fraud	99.4	99.7	99.6
Accuracy	99.7	99.7	99.7
Macro AVG	99.6	99.7	99.7
Weighted AVG	99.7	99.7	99.7

This confusion matrix provides a granular view of the model's performance. As detailed in Figure 5, the model correctly classified 19654 instances as Not Fraud and 13017 instances as Fraud. However, there were 72 instances where the model incorrectly classified 'Not Fraud' instances as 'Fraud' (False Positives) and 33 instances where 'Fraud' was incorrectly classified as 'Not Fraud' (False Negatives) [14].

The minimal number of false positives and false negatives suggests that the model has high precision and recall, confirming the metrics observed in the classification report. This further emphasizes the robust performance of the model in fraud detection, underscoring its potential for reliable use in real-world applications.

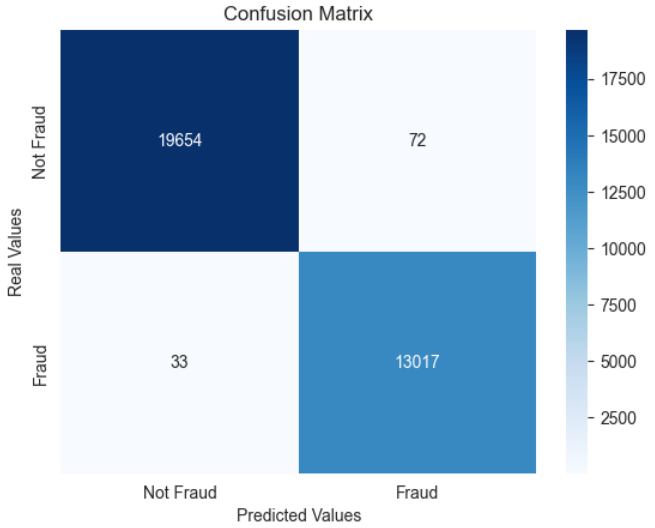


Figure 5. Confusion matrix of the model evaluation over the test set.

The AUC (Area Under the Curve) score and the ROC (Receiver Operating Characteristic) curve are critical metrics for evaluating the performance of a binary classification model. The AUC score for this model is 0.997, which is remarkably close to 1. This indicates that the model has a high measure of separability and is highly capable of distinguishing between positive and negative classes. The ROC curve, which is a plot of the true positive rate against the false positive rate, provides a visual representation of the model's performance across all thresholds. The curve for this model appears to be close to the ideal top-left corner of the plot, suggesting a high true positive rate and a low false positive rate. The previous description is detailed in Figure 6 [15].

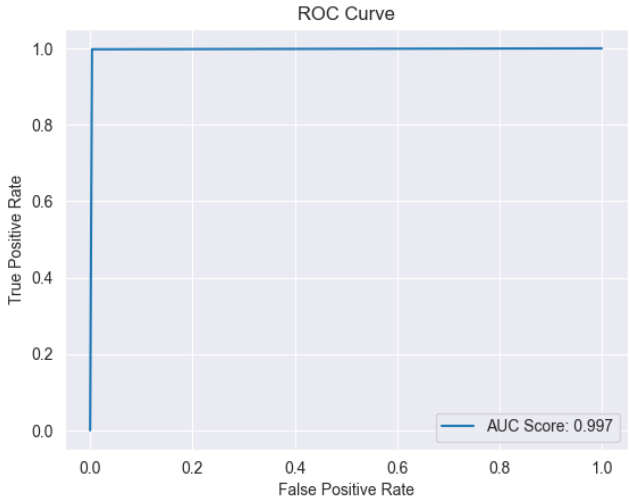


Figure 6. ROC Curve of the model evaluation over the test set

The Dense Neural Network model proposed in this research for fraud detection represents a significant contribution to the financial sector and cyber-security. Its high performance, as evidenced by the AUC score and the confusion matrix, demonstrates its robust ability to accurately distinguish between fraudulent and non-fraudulent transactions. This accuracy is paramount in the financial sector, where the timely detection and prevention of fraudulent activities can result in substantial cost savings and enhance the security of financial transactions.

Table 2. Comparative analysis of model performance in the literature.

Model	Accuracy (%)	Precision (%)	Recall (%)	Learning Type
Varmedja et al [4]: LR	97.46	58.82	91.84	Supervised
Varmedja et al [4]: NB	99.23	16.17	82.65	Supervised
Varmedja et al [4]: RF	99.96	96.38	81.63	Supervised
Varmedja et al [4]: MLP	99.93	79.21	81.63	Supervised
Xuan et al [5]	98.67	32.68	59.62	Supervised
Dornadula et al [6] RF	99.9	**99.9**	-	Supervised
Jhon & Naaz [7] LOF	97	-	-	Unsupervised
Jhon & Naaz [7] IF	71	-	-	Unsupervised
Zadafiya et al [8] IF	**100**	65	64	Unsupervised
Zadafiya et al [8] IF	**100**	51	51	Unsupervised
Proposed DNN Model	99.7	99.6	**99.7**	Supervised

The comparison delineated in Table 2 involves models that utilize distinct datasets for their training and evaluation processes. In this table, values are emphasized in bold to

indicate superior performance, however, it should be noted that in the studies conducted by Dornadula et al. [6], the recall metrics were not evaluated, similarly, in the research undertaken by John and Naaz [7], neither precision nor recall was assessed. Despite other models outperforming the proposed model in certain metrics, the proposed model generally exhibits superior performance across all metrics in the comparison. The model from the Dornadula et al [6] study emerges as the most competitive against the proposed model in terms of accuracy and precision, although a comparison of recall metrics is not possible.

In the realm of cybersecurity, the model's ability to detect anomalies and classify transactions with high precision contributes to the strengthening of security protocols. By identifying potential threats, it aids in the proactive mitigation of cyber risks, thereby enhancing the overall security posture of financial institutions. Furthermore, the adaptability of the model allows for its integration into the existing infrastructure of any financial institution or bank. By training the model on institution-specific data, it can be tailored to detect fraud patterns unique to the institution, thereby increasing its effectiveness.

The model can be integrated as an AI monitoring component into a real-time Anti-Money Laundering (AML) system. Utilizing technologies such as Kafka and PySpark for data sourcing and streaming, the model can analyze and classify transactions in real time. This not only allows for immediate detection and response to fraudulent activities but also enables continuous learning and adaptation to evolving fraud patterns.

IV. Conclusions

While acknowledging that an Anti-Money Laundering system consists of more components than just AI monitoring, the focus of this paper has been on the development and evaluation of an AI model for transaction fraud detection based on deep learning.

The proposed model, a dense neural network, has exhibited exceptional performance, with precision, recall metrics, and an Area Under the Curve score all approximating 99.7%. The model's robustness is further underscored by its use of a class balance method, ensuring an unbiased generalization capability. Comparative analysis with extant models reveals that the proposed model generally outperforms on key performance metrics. Its architecture is designed for seamless adaptability and integration into pre-existing financial systems, thereby bolstering the security of financial transactions and the integrity of the financial system at large. This research represents an important milestone and lays a solid foundation for future research and implementation of an anti-money laundering system specifically tailored to the Cuban banking sector, which currently lacks a proactive system of this type.

References

[1] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI). http://dx.doi.org/10.1109/ICCNI.2017.8123782

[2] Narayan, A., Kumar, S. D. M., & Chacko, A. M. (2023). A Review of Financial Fraud Detection in E-Commerce Using Machine Learning. First Online: 24 February 2023. 346 Accesses. Conference paper. http://dx.doi.org/10.1007/978-981-19-7524-0_21

[3] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., et al. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. Applied Sciences, 12(19), 9637. MDPI AG. http://dx.doi.org/10.3390/app12199637

[4] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). https://doi.org/10.1109/infoteh.2019.8717766

[5] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). https://doi.org/10.1109/icnsc.2018.8361343

[6] Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

[7] John, H., & Naaz, S. (2019). Credit card fraud detection using local outlier factor and isolation forest. Int. J. Comput. Sci. Eng, 7(4), 1060-1064.

[8] Zadafiya, N., Karasariya, J., Kanani, P., & Nayak, A. (2022). Detecting Credit Card Frauds Using Isolation Forest And Local Outlier Factor-Analytical Insights. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1588-1594). IEEE

[9] Narayanan R, D. (2021). Credit Card Fraud. Kaggle. Available at: https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

[10] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. https://doi.org/10.48550/arXiv.1207.0580

[11] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). pmlr. https://doi.org/10.48550/arXiv.1502.03167

[12] Keras. Keras API Reference. Available online: https://keras.io/api. Last Accessed 28/3/2023

[13] Scikit-learn. (2023). Classification Report. Available at https://scikit-learn.org/stable.

[14] Ting, K.M. (2011). Confusion Matrix. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_157

[15] Hoo, Z.H., Candlish, J., Teare, D., (2017). What is a ROC curve?. Emergency Medicine Journal 34, 357–359.. https://doi.org/10.1136/emermed-2017-206735