Medical Data Analysis of Polycystic Ovarian Disease using Deep Learning

V. Shoba, PhD
Assistant Professor, Department of Computer Applications & Technology,
SRM Arts and Science College, Kattankulathur-603203

D. Bhuvaneswari, PhD
Assistant Professor, Department of Computer Scicene,
SRM Arts and Science College, Kattankulathur-603203

ABSTRACT
Recent Events, Polycystic Ovarian Disease (PCOD) is very important in the realm of women's lives. PCOD is mostly caused by a hormonal imbalance and inherited predisposition. Each month, the two ovaries alternately release mature, ready-to-fertilize eggs in a typical menstrual cycle. For the prepossessing, the PCOD dataset is downloaded from the Kaggle repository as a.csv file type. In order to input into the prediction, pre-processing involves removing unnecessary data and filling in missing values. Currently, the analysis utilizes three deep learning algorithms for disease prediction: Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN) .The forecast takes into consideration the patient's age, height, weight, and particular characteristics like FSH, LH, endometrium thickness, II beta and I beta HCG, as well as whether or not the patient is pregnant. Pre-processed datasets are classified and algorithm accuracy is assessed using Deep Learning Models like DNN, RNN, and CNN. The performance of each algorithm is assessed by comparing classification metrics such as precision, recall, and f-measure values. Among them, DNN outperforms the others. Other categorization techniques were then employed to find enormous amounts of data.

Keywords
Polycystic Ovarian Syndrome, DNN, RNN, CNN

1. INTRODUCTION
Collaboration between humans and innovation could lead to improved health care services. Artificial intelligence's deep learning branch enables an algorithm to learn and develop without explicit programming. The main objective is to provide novel methods for deep learning that facilitate the utilisation of specific datasets for open-source research and analysis. Deep learning applications have brought about significant advancements, particularly in the healthcare sector, where they are used for tasks such as diagnosis, image recognition, data analysis, prediction, and more. Polycystic ovarian syndrome (PCOS), an endocrine disorder, commonly affects women during their teenage years. Leventhal and Stein told it for the first time in 1935. Hormone levels in women having polycystic ovarian syndrome are noticeably out of whack. It results in serious health issues like irregular menstruation cycles and trouble getting pregnant.

Women with PCOS are at an increased risk for conditions such as hypertension, cardiovascular disease, type 2 diabetes, obesity, gynecological cancers, high-risk pregnancies, and diabetes mellitus. Symptoms of PCOS include acne, high blood pressure, irregular menstrual cycles, weight gain, elevated androgen levels, and hormonal imbalances, among others. PCOS is considered a leading cause of infertility as it disrupts follicular development, impairing ovarian maturation. A recent study has highlighted a significant risk of first-trimester miscarriage in individuals with PCOS. This condition affects approximately 12-21% of women of reproductive age, with 70% of cases remaining undiagnosed. PCOS can be managed through doctor-prescribed medications and lifestyle modifications. Treatment options include birth control pills, diabetes medications, anti-androgen drugs, fertility evaluations, and ultrasounds. Diagnosis is typically made by ruling out unrelated symptoms or test results, as PCOS is often linked to a complex and poorly understood pathomechanism[1].

Due to the wide range of symptoms, many unnecessary radiological imaging tests and clinical investigations are often required [2]. The processes involved in conception, ovulation, and fetal development in a woman's womb are heavily influenced by hormones that must remain in balance within the female reproductive system. These essential hormones include estrogen, progesterone, luteinizing hormone (LH), and follicle-stimulating hormone (FSH). LH and FSH are produced by the pituitary gland, while the ovaries generate progesterone and estrogen. Proper functioning of the female reproductive system relies on both progesterone and estrogen. Women with polycystic ovarian syndrome (PCOS) face a range of challenges, including sleep apnea, infertility, uterine bleeding, elevated lipids and cholesterol, non-alcoholic fatty liver disease, anxiety, depression, hypertension, metabolic syndrome, miscarriages, and an increased risk of cardiovascular issues. Between 30% and 40% of women with PCOS experience symptoms such as amenorrhea, abdominal obesity, enlarged breasts before menstruation, and excessive or unwanted facial or body hair. They may also suffer from neuralgic pain, hysteria, vulvar and vaginal discomfort, and ovarian cysts[3].

The remaining study was arranged as follows. The literature review is covered in Section II, the problem formulation for this research project is covered in Section III, and Recurrent neural networks (the RNN), convolutional neural networks (the CNN), and deep neural networks (the DNN) —for determining correctness are covered in Section IV. In section V, the discussion and outcomes of the experiment are shown. Section VI contains the innovative information that brings this study to a close.

2. LITERATURE SURVEY
Bharati S. et al. predicted PCOS using Kaggle ML models. For instance, the authors used the PCOS dataset and a univariate feature selection (UFS) approach to apply gradient boosting, RF, LR, with a hybrid RFLR model that combined RF and LR. To train and test the models, they divided the dataset utilising holdout and cross-validation techniques. According to the results, RFLR with UFS worked the best. Principal Component Analysis (PCA) was used by the authors of [5] in order to decrease the number of attributes. To predict PCOS, they used

NB, KNN, LR, RF, as well as SVM together with a few well-selected characteristics. According to the findings, RF had the best accuracy. To pick a subset of attributes from the database, the authors of [6] employed correlation feature selection techniques. Based on correlation levels, they used SVM, LR, RF, DT, KNN, QDA, LDA, GB, AdaBoost (AB), XGBoost (XB), and CatBoost to determine which model was the best. Based on the findings, RF was the most effective model.

The authors of [7] evaluated a number of models, including CNN, ANN, SVM, DT, then KNN, and used feature selection techniques to diagnose PCOS. The best-performing model was made by RF. Utilising Pearson correlation, the best features in [8] were assessed. When evaluating the accuracy rate of their SVM, the applied SVM, RF, along with XG boost multi-layer perceptron using chosen features provides the highest accuracy rate. To minimise the number of features, the researchers of [9] presented a combination feature selection method that uses wrappers and filters. To predict PCOS, they also used several machine learning models with distinct characteristics. Most accurate model was SVM.

The goal of Palak et al.'s work is to develop an automated technique for PCOS screening that is based on metabolic and clinical markers. The study technique uses logistic and Bayesian regression to categorise characteristics. Out of the two models that were assessed, the Bayesian classifier had the highest accuracy (93.93%) and was the best-constructed model [10]. With Denny and associates [11], The goal is to circumvent the time and money associated with other clinical diagnostic procedures, such as ovarian screening. The PCA is used in the research design to translate PCOS characteristics using machine learning techniques like KNN, SVM, RF, and others. Random Forest produced the most efficient and accurate approach to PCOS identification, having an accuracy of 0.89.

Pijush et al. [15] discuss the early detection and management of this illness. SMOTE was combined with five other algorithms—Random Forest, Decision Tree, Support Vector Machine, KNN, and Logistic Regression—to detect PCOS at an early stage. The most accurate model produced the following outcomes: 98% recall, 98% precision, 95.6% AUROC, 97.11 training time, and 0.010 seconds F1 score. A probabilistic method was employed by Khan Inan et al. [16] to determine the statistically relevant traits connected to PCOS cases. Important features were found using the Chi-Square test, the ENN, ANOVA, and SMOTE tests. XG Boost, SVM, KNN, NB, MLP, RF, and AdaB classifiers were among the ones employed. G Boost outperformed all other classifiers, returning 0.98 and having an accuracy of 0.96.

The LeBERT sentiment classification approach, which combines a sentiment lexicon, N-grams, BERT, and CNN, is described by Mutinda et al. in [17] and gets over these limitations. The model vectorizes words from a subset of the input text using sentiment lexicon, N-grams, & BERT. CNN, a deep neural network classifier, generates an output sentiment category following feature mapping. Three publicly available datasets are used to assess the proposed method: Yelp restaurant reviews, Amazon retail reviews, and IMBD movie reviews. Text mining & text processing techniques were employed by S. Vijayarani et al. [18] to find knowledge in text material submitted by social media users. They talked about the TF/IDF approach, stop word removal, stemming, and the first step in text preprocessing. The stemming algorithms of each group were also investigated, and they revealed the benefits and drawbacks of various phases. This study illustrated each stage in text preparation that was required for sentiment analysis or text mining.

3. PROBLEM STATEMENT

This section discusses the research work's problem definition. The organisation of various types of unstructured data in the medical record is a major difficulty in medical data mining jobs. Accurate illness diagnosis using medical data necessitates an understanding of the patterns and important terms in a patient's medical history, which might vary greatly. The dataset, which contains redundant data, missing data, and irrelevant features, is preprocessed, and the cleaned PCOS dataset is used in the prediction process using DNN, RNN, and CNN algorithms to predict whether or not the patients are impacted. The PCOD dataset is downloaded as a.csv file from the Kaggle source. The dataset's filename is PCOS_data.csv.

Sl. No	Patient File No.	Age (yrs)	Weight (K	Height(Cn	BMI	Blood Gro	Pulse rate	RR (breath	Hb(g/dl)	Cycle(R/I)	Cycle leng	Marraige S	Pregnant(	No. of abc	I	beta-H	II	beta-H	FSH(mIU/	LH(mIU/m
1	1	28	44.6	152	19.3	15	78	22	10.48	2	5	7	0	0		1.99		1.99	7.95	3.68
2	2	36	65	161.5	24.9	15	74	20	11.7	2	5	11	1	0		60.8		1.99	6.73	1.09
3	3	33	68.8	165	25.3	11	72	18	11.8	2	5	10	1	0		494.08		494.08	5.54	0.88
4	4	37	65	148	29.7	13	72	20	12	2	5	4	0	0		1.99		1.99	8.06	2.36
5	5	25	52	161	20.1	11	72	18	10	2	5	1	1	0		801.45		801.45	3.98	0.9
6	6	36	74.1	165	27.2	15	78	28	11.2	2	5	8	1	0		237.97		1.99	3.24	1.07
7	7	34	64	156	26.3	11	72	18	10.9	2	5	2	0	0		1.99		1.99	2.85	0.31
8	8	33	58.5	159	23.1	13	72	20	11	2	5	13	1	2		100.51		100.51	4.86	3.07
9	9	32	40	158	16	11	72	18	11.8	2	5	8	0	1		1.99		1.99	3.76	3.02
10	10	36	52	150	23.1	15	80	20	10	4	2	4	0	0		1.99		1.99	2.8	1.51
11	11	20	71	163	26.7	15	80	20	10	2	5	4	1	2		158.51		158.51	4.89	2.02
12	12	26	49	160	19.1	13	72	20	9.5	2	5	3	0	1		1.99		1.99	4.09	1.47
13	13	25	74	152	32	17	72	18	11.7	4	2	7	1	0		1214.23		1214.23	2	1.51
14	14	38	50	152	21.6	13	74	20	12.1	2	5	15	0	0		1.99		1.99	4.84	0.71
15	15	34	57.3	162	21.8	13	74	22	11.7	2	5	9	0	0		1.99		1.99	7.45	3.71
16	16	38	80.5	154	33.9	13	78	22	11.4	2	5	20	0	0		1.99		1.99	9.51	2.51
17	17	29	43	148	19.6	13	80	20	11.1	2	5	2	1	0		8104.21		91.55	2.02	0.65
18	18	36	69.2	160	27	13	72	18	10.8	2	5	7	0	0		1.99		1.99	4.86	2.96
19	19	31	52.4	159	20.7	17	72	18	12.7	2	5	7	0	0		1.99		1.99	6.05	1.05
20	20	30	85	165	31.2	16	72	18	12.5	4	7	7	0	0		23.58		1.99	1.89	0.81
21	21	25	64	156	26.3	11	70	18	11.2	2	6	6	0	0		1.99		1.99	2.82	1.3

Figure 1. Sample Dataset

This research work comprises over 5000 records and analysed the records of 681 patients. The dataset is in.CSV format and was created for this study. The dataset has 43 attributes, and 681 records are considered. The individuals in this study are either affected by PCOS or not. Figure 1 depicts a sample dataset. The characteristics differ from one sufferer to the next.

4. RESULTS

This section provides a full report on the results obtained by the three existing PCOS Disease algorithms, DNN, RNN, and CNN, which are programmed in the Python computer language. Table 1 displays the number of patients affected by PCOS Disease, and Figure 4 depicts the dataset's graphical form.

TABLE 1: Number of Patients affected by PCOS Disease

PCOS	Number of Patients
Affected	434
Not Affected	247
Total	681

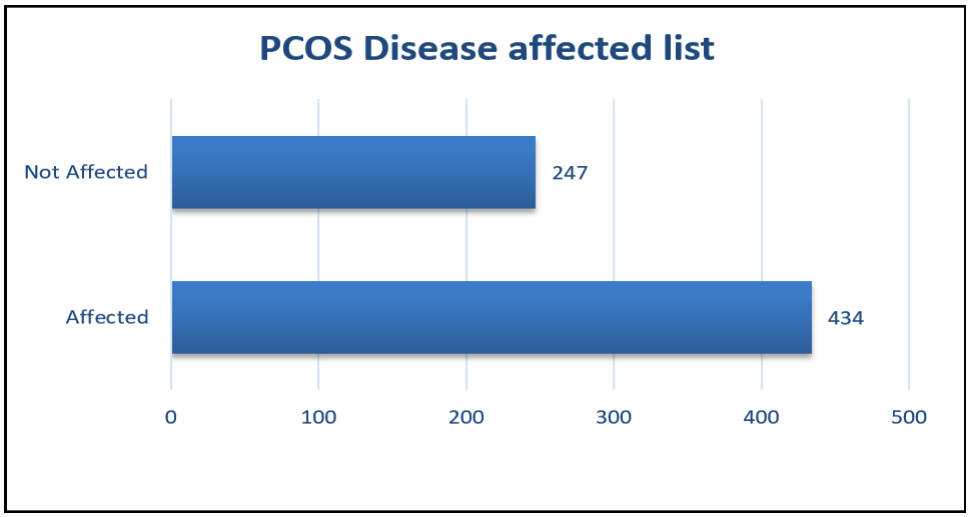

Figure 3. Number of Patients Affected by PCOS Disease

The three approaches are evaluated for effectiveness using the following metrics: f-measure, recall, accuracy, and precision. The precision, recall, & f-measure rates for each of the three algorithms are shown in Figure 4.

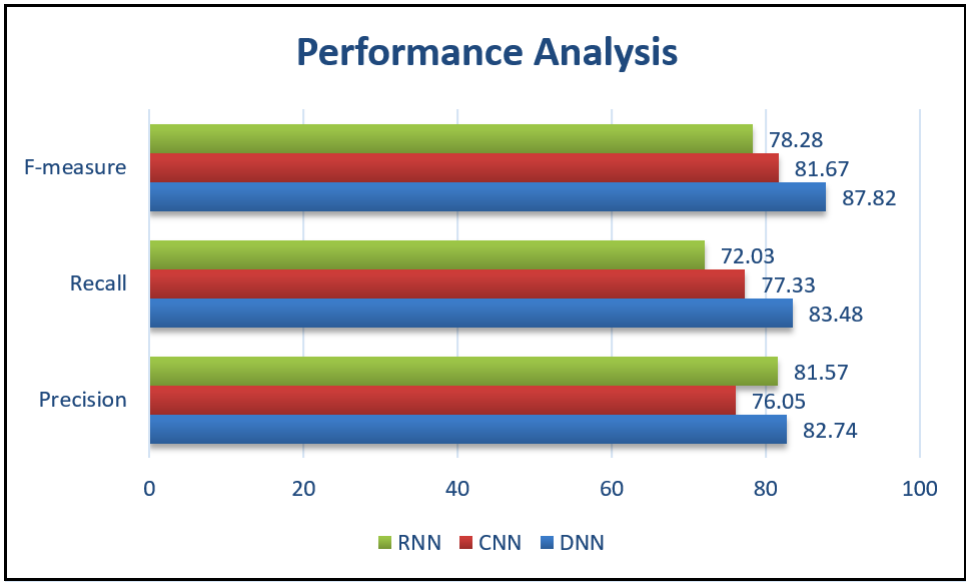

Figure 4. Performance Analysis

The DNN algorithm achieved a precision score of 82.74%, recall score of 83.48%, & F-measure score of 87.82%. Figure 4 shows the performance evaluation for all three approaches. The CNN algorithm yields an 81.67% f-measure value, 77.33% recall, and 76.05% precision. The RNN Algorithm obtains an F-measure score of 78.28%, a precision score of 81.57%, and a recall score of 72.03%. It is clear from the data that the algorithms DNN algorithm performs better than the other two methods that are currently in use.

TABLE 2: Performance Analysis of DNN, CNN and RNN

Algorithms	Precision	Recall	F-measure
DNN	82.74	83.48	87.82
CNN	76.05	77.33	81.67
RNN	81.57	72.03	78.28

Table 2 displays the performance analysis, and Table 3 displays the time and memory usage for each algorithm.

TABLE 3: Average Computational Time and Memory Utilization of Algorithms

Algorithms	Execution time (ms)	Memory utilization (bits)
DNN	2725	125579
CNN	3242	179366
RNN	4011	237399

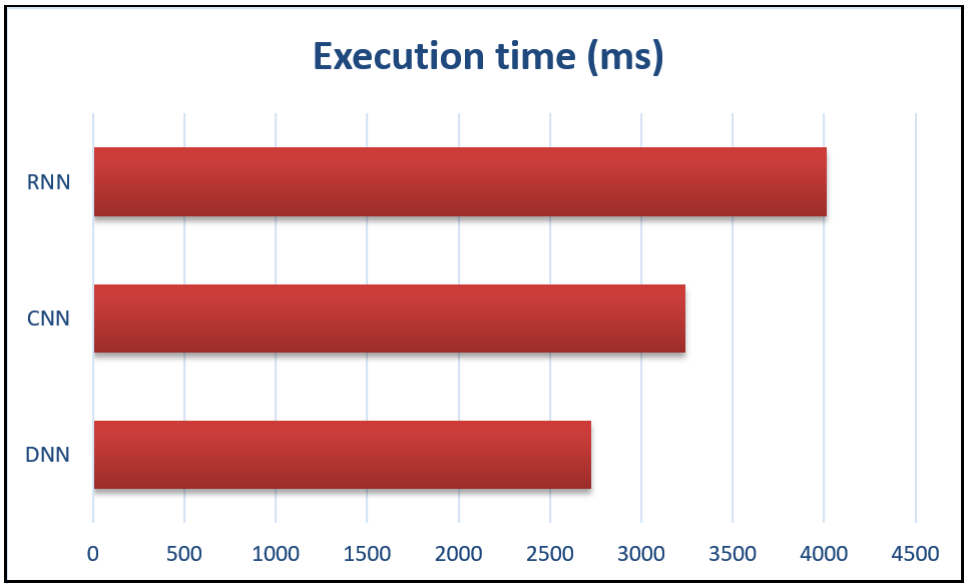

Figure 5. Run Time

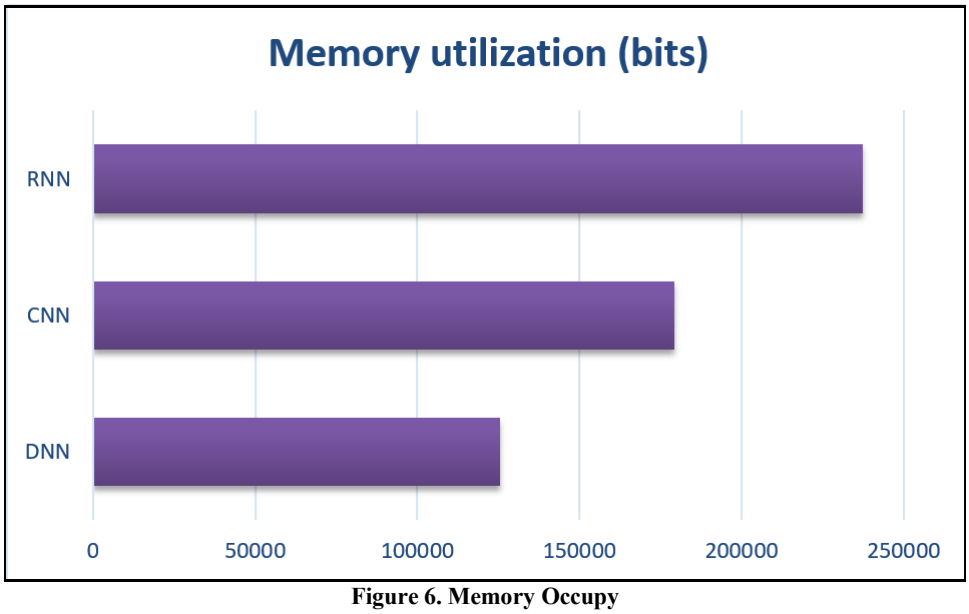

Figure 6. Memory Occupy

A graphical representation of the three different classification algorithms' execution times for the datasets they produced is shown in Figure 5. Figure 6 displays a graphical depiction of the amount of memory space used by the final set of data for each of the three categorization techniques. Figure 6 shows that compared to RNN and CNN approaches, the DNN methodology computes much more quickly. For the given dataset, Figure 6 illustrates that the DNN algorithm uses less memory than the RNN & CNN algorithms.

5. CONCLUSION

Predicting the optimal classification method for every dataset is generally impossible. However, the performance of each classification system varies depending on the dataset used for analysis. In real-world applications, classification algorithms play a crucial role in analyzing different types of data. This study utilizes the PCOS Disease Dataset, which provides two possible outcomes: whether a patient is affected by PCOS or not. The results indicate that three-fourths of the patients in the dataset are impacted by PCOS. These findings are derived from various features within the dataset, and the performance metrics of all three classification methods are evaluated using precision, recall, and F-measure. Among the methods tested, the Deep Neural Network (DNN) technique achieves the highest precision, recall, and F-measure values. The study confirms that the DNN approach demonstrates superior accuracy in predicting PCOS compared to Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Additionally, DNN effectively identifies different entity types across various datasets without requiring explicit annotations. Future research presents opportunities to incorporate additional prediction algorithms to further enhance forecasting accuracy.

6. REFERENCES

[1] Aroni Saha Prapty and Tanzim Tamanna Shitu, “An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome”, *23rd International Conference on Computer and Information Technology (ICCIT)*, doi:10.1109/ ICCIT51783.2020.9392666, pp. 1–5, 2020.

[2] Amsy Denny, Anita Raj, Ashi Ashok, C Maneesh Ram, and Remya George, “i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques”, *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON),* doi:10.1109/ TENCON.2019.8929674, 2019, pp. 673–678.

[3] Bharati S., Podder P., Mondal M.R.H., “Diagnosis of polycystic ovary syndrome using machine learning algorithms”, *Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, 5–7 June 2020, pp. 1486–1489.

[4] Tiwari S., Kane L., Koundal D., Jain A., Alhudhaif A., Polat K., Zaguia A., Alenezi F., Althubiti S.A., “SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning”, *Expert Syst. Appl*., doi: 10.1016/ j.eswa.2022.117592, 2022.

[5] Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis”, *Proceedings of Academia-Industry Consortium for Data Science: AICDS*

[6] Priyanka R. Lele, Anuradha D. Thakare, “Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures*”, International Journal of Engineering Research & Technology (IJERT)*, Volume 9(3), March-2020, ISSN: 2278-0181.

[7] Namrata Tanwani, “Detecting PCOS using Machine Learning”, *IJMTES | International Journal of Modern Trends in Engineering and Science*, Volume 7(1), 2020, ISSN: 2348-3121.

[8] Pijush Dutta, Shobhandeb Paul, Madhurima Majum-der, “An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS”, *Research Square*, November 8th, 2021.

[9] Mutinda, James, Waweru Mwangi, and George Okeyo, "Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network", *Applied Sciences*, Volume 13(3), pp.1445, 2023.

[10] T. Vaiyapuri, A. K. Dutta, and I. S. Punithavathi, “Intelligent deep-learning-enabled decision-making medical system for pancreatic tumor classification on ct images”, *Healthcare*, Volume 10(4), pp. 677, 2022.

[11] Altman, E.I., 1968. “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, *The journal of finance*, Volume 23(4), pp. 589-609.

[12] Hamori, S., Kawai, M., Kume, T., Murakami, Y. and Watanabe, C., “Ensemble Learning or Deep Learning? Application to Default Risk Analysis”, *Journal of Risk and Financial Management*, Volume 11(1), pp. 12, 2018.

[13] Hinton, G.E. and Salakhutdinov, R.R., “Reducing the dimensionality of data with neural networks. Science”, 2006, pp. 504-507.

[14] Larochelle, H., Mandel, M., Pascanu, R. and Bengio, Y., “Learning algorithms for the classification restricted Boltzmann machine”, *Journal of Machine Learning Research*, Volume 13, 2012, pp. 643-669.

[15] Mazumder, R., Hastie, T. and Tibshirani, R., “Spectral regularization algorithms for learning large incomplete matrices”, *Journal of machine learning research*, Volume 11, 2010, pp. 2287-2322.