Revolutionizing Medical Diagnosis with Convolutional Neural Network: A Data-Driven Solution to Improve Accuracy in Disease Detection

Mohammad Rifaq[*], Daud Khan[†]

Abstract

Currently, the global mortality rate is on the rise due to the increasing prevalence of various diseases. This surge in deaths is largely attributed to the growing number of patients suffering from major health conditions. Unfortunately, many patients are often misled by medical practitioners, and in some cases, these physicians lack the necessary expertise to accurately diagnose specific diseases. This presents a significant challenge in healthcare today. Timely and accurate diagnosis of diseases remains a critical issue due to a shortage of specialists and a lack of experience in managing similar cases. Traditional manual diagnostic systems for diseases are often unreliable due to uncertainties in clinical data and medical knowledge. In this research, a deep learning algorithm is developed, specifically a Convolutional Neural Network (CNN), to detect diseases. A datasets from the UCI Machine Learning Repository is utilized, applying machine learning techniques to diagnose these conditions without requiring the assistance of medical professionals. Additionally, the proposed model can predict whether an individual is at risk of developing a particular disease in the future based on key factors associated with the disease. The medical datasets used in the proposed study include Breast Cancer, Diabetes, Heart Disease, and Liver Disorders from UCI. The data underwent preprocessing before being input into the CNN. The accuracy of the proposed model is compared with previous research, demonstrating its effectiveness and superior training.

***Keywords**:* Deep Learning; Medical Diseases; Disease Diagnosis; Convolutional Neural Network.

Introduction

In medical diagnosis, machine learning algorithms analyze specific datasets to identify diseases. This paper presents the application of different machine learning systems on various medical datasets (specifically related to breast cancer, diabetes, heart disease and liver disorder). These methods help the physicians in accurate diagnosis, which is essential for effective treatment, and contribute useful predictions to improve overall diagnostic accuracy (Sujith et al., 2023). The worldwide mortality rate is growing day by day with most of the world now suffering from these diseases. These diseases have been a rapidly growing epidemic

[*] Department of Computer Science , Iqra National University, Peshawar 25100, Pakistan, m.rifaq@yahoo.com

[†] Corresponding Author: Department of Computer Science , Iqra National University, Peshawar 25100, Pakistan, daud.khan@inu.edu.pk

according to the World Health Organization (WHO). Studies show that misdiagnosis of these diseases and late diagnosis cause a stake in the prognosis. Even for the medical fraternity timely and correct diagnosing abbreviation always remains a major challenge since they are too short with specialists or past cases (Wee et al., 2024). Traditional manual diagnostic methods have been proved to be not credible since those studies by nature are categorical and full of inevitable ambiguity with regards to clinical or medical data, stressing the need for more reliable and alike well-performed approaches through machine learning. Historically, diagnosis disease is expensive and heavily dependent on expensive medical skills plus data. Instead, A machine learning algorithm can be trained on past disease datasets to diagnose conditions in a manner similar to the diagnostic processes performed by physicians today (Sai Shekhar et al., 2020).

One of the important functions of these techniques, especially classification methods, is in detecting diseases systematically. Here, patient records have been traditionally used for diagnosis by medical professionals but using an automated diagnostic system can cost savings on the valuable time and money. The rise of the need for applications like deep learning that can diagnose with reasonable conditions, expert systems that allow rapid and precise diagnosis of complex diseases from extensive data, is bound to happen in future so as to treat potential early stages of some major causes with absolute certainty. In other words, the field of computerized expert systems has developed an approach to diagnosing diseases from datasets for computational analysis (Reddy et al., 2022). The system itself is an example of Artificial Intelligence (AI) that as a deep learning model, it mimics the way a human brain works through an Artificial Neural Network (ANN). The ANN consists of three layers, namely Input layer → hidden layer → output layer → called nodes available for processing data (Patil et al., 2022). The data gets fed in the first layer, goes to the hidden layer which holds neurons and results come up in the output layer. In the processing layer, summation and transfer functions are applied to get final output. The task is to classify (diagnose) high-risk diseases with significant mortality rates by using few deep learning methodologies, especially Convolutional Neural Network (CNN). These findings are compared with the results of previous studies in order to increase the accuracy and performance of diagnosis. This study is based on the medical research database used for the UCI Machine Learning Repository (Nabeel et al.,2021).

UCI Machine Learning Repository is a popular source, and they are widely available and preprocessed, making the UCI Machine Learning Repository a convenience source for benchmarking models on a wide range of datasets. This study applies it to small but well-documented datasets (Breast Cancer, Diabetes, Heart Disease, and Liver Disorders) to ascertain machine learning feasibility before applying it to more complicated real-life data (Khan et al., 2021).

This research focuses on the application of complicated algorithms in many domains, including medical diagnosis. It helps develop expert systems for pre-alarm and detection of diseases like breast cancer, diabetes, heart disease, and liver diseases with light-based modalities for automated diagnosis without a physician. The model improves diagnostic software for automated systems, as well as for real-time classification tasks such as pattern recognition, fraud detection, and communications. Its goal is to enhance both diagnosis and prediction quality to achieve early detection of diseases.

Literature Review

Traditional Chinese Medicine expert systems are central to automatic diagnosis, researched over decades. Experts manually collect and structure data in normalized forms. Later, various researchers work on this data, but their results are not good because they don't apply it effectively to clinical datasets and use poorly structured forms (Wang et al., 2021). This dataset is unstructured and not validated by medical boards. The author introduces automatic diagnosis techniques to investigate the Traditional Chinese Medicine dataset using two machine learning methods: Naive Bayes (based on probability) and Support Vector Machine (using hyperplanes for classification). Both methods are compared for performance, and the better one is identified. Breast cancer, a common and serious disease in women, results from abnormal cell growth, either cancerous or non-cancerous, treatable with proper care. Breast cancer diagnosis traditionally depends more on the expertise of the doctor involved; nevertheless, recent systems use the clinical data of the patient instead and can identify the classification of the patient more accurately. Three machine learning methods are applied to the UCI breast cancer database in (Hota et al., 2023). ANN, statistical methods, and Decision Trees. Likewise Bahramirad et al. (2021) give a review of liver disease data using eleven machine learning algorithms and determine the most precise algorithm. They allow for the creation of expert systems for disease diagnosis, prognosis, and treatment. Results also improve using WEKA data mining techniques in (Ramzan et al., 2021).

Thomas et al. (2022) utilize different types of machine learning classification techniques to diagnose heart disease, and the researcher uses basic medical parameters of humans to predict disease. Sabariah et al. (2022) have combined different data mining methods to make a classification framework for early diagnosing of disease and early detection of type 2 diabetes patients. Two methods are used in this paper; CART (Classification and Regression Trees) is used for complex datasets, implementing this method. Bendi et al. (2020) use the Multiple Classifier Systems (MCS) Rotation Forest data mining technique to classify cancer disease patients using microscopic or microarray medical databases from the medical dataset repository. According to the researchers, the Rotation Forest technique is the best for classifying cancer disease, so they apply the Random Forest data mining technique on a database based on microarray values for classification purposes. Rathinaeaswari et al. (2023) propose a feature selection-based ensemble classification methodology in the ensemble system for data mining purposes with a high accuracy rate. K Subsections are applied to the feature set, and each subsection is distributed to Principal Component Analysis (PCA).

To reserve useful database information, all basic components are used. Thangavel et al. (2024) use the ensemble technique for better performance of a base classifier. The dataset of three diseases is taken and applied the Rotation Forest model to those datasets. As different algorithms of machine learning techniques are used in this research, results show that the Random Forest model gives better and best results with high accuracy. For diagnosing liver disease, the researchers use many machine learning methodologies on patient datasets for the classification of disease. ANN, which behaves like the human brain's operation of biological methodology. Different machine learning methodologies are applied to two different datasets of liver disease on specific attributes and perform analysis studies on them. Performance accuracy is best when applied K-Nearest Neighbor (KNN) for an Indian patient's liver medical dataset. KNN classifies the closed occurrence values (Allenki et al., 2024). Jaisinghani et al. (2023) use WEKA for the implementation of machine learning techniques. It is very helpful for classification and is based on built-in functionality for data mining of desired datasets for classification. Liver disease is a very common disease nowadays around the world, for diagnosing liver disease the author of this research uses the Bayesian classification algorithm, which consists of two types of machine learning methodologies: Boosting and Bagging respectively. Three machine learning methodologies are used: Naïve Bayes works on a probability basis, to increase accuracy the Bagging methodology is used, and the last one is an ensemble model technique. Boosting works the same way as the Bagging technique.

Decision Tree techniques are the best machine learning technique for classification of data. Kawarkhe et al. (2024) show that the best classification methodology is Decision Tree algorithms, four types of

Decision Tree techniques are used in this research work. Ensemble techniques perform better than individual classification methodologies. Better classification of liver disease is obtained by modification of the Rotation Forest technique, which is carried out by two classification methodologies: selection-based and feature-based.

For classification purposes, different selection-based algorithms are performed and compared with each other. On the UCI dataset of liver disease, the modified algorithm of Random Forest is applied; on the other hand, KNN and feature-based algorithms are applied on the Indian liver dataset. This research shows that the ensemble model gives the best accuracy with the lowest fault factor rather than any single algorithm. Different diseases' datasets are available on the UCI repository website. 497 datasets of different diseases are on this website for performing machine learning algorithms. Machine learning algorithms are performed on these datasets. This dataset is very useful for medical science research. Most of the researchers use datasets of diseases from this platform which is known as UCI Repository for Machine Learning Techniques. A comparative study of different machine learning algorithms is evaluated in this research (Vellela et al., 2023). The main factor in the increase of liver disease is the lower prediction of disease occurrence in human beings. Performance accuracy of liver disease for diagnosing is obtained by using CBR and CART techniques to make an intelligent model for diagnosing.

Classification and Regression Tree give 92.94% accuracy and Case-Based Reasoning method gives 90%, results show that CART is best from CBR. Data enters the neural network by using the input layer, and the connection between both input and output is the hidden layer which contains neurons. WEKA provides full access to researchers to perform all types of machine learning algorithms without any trouble with built-in functionalities for all kinds of datasets for classification (Ahmad et al., 2024).

Bendi et al. (2020) leverage machine learning techniques to tackle medical challenges, developing AI systems capable of tasks such as predicting breast cancer risk. Their work utilizes data from the UCI database, incorporating variables like anthropometric measurements and blood test results. They also explore machine learning applications in heart disease diagnosis, emphasizing the intricacies of these systems and their dependence on comprehensive patient data. Likewise, Hashem et al. (2023) assess machine learning models for forecasting advanced liver fibrosis in chronic hepatitis C patients. Their analysis involves clinical records and serum biomarker data from 39,567 individuals, classifying them into groups with either mild/moderate or advanced fibrosis.

To evaluate the performance of these models, ROC curve analysis is proposed. Devika et al. (2024) conduct a comparative study for the classification of chronic kidney disease prediction by using Naïve Bayes, KNN, and Random Forest. Similarly, Sisodia et al. (2024) conduct a study to know the performance of individual and ensemble learners for Chronic Kidney Disease (CKD) prediction. They predict CKD by using individual and individual classifier three different classifiers: Naïve Bayes, Minimal Sequential Optimization, J48, and for ensemble classifier Random Forest, Bagging, AdaBoost are used for prediction of CKD. Results are evaluated by using recall, F-measure ROC, and accuracy performance. Results show that individual learner J48 and ensemble classifier from random forest perform best than other classifiers.

Thenmozhi et al. (2023) conduct a study for the prediction of heart disease by using different Decision Tree techniques. Similarly, Haseeb et al. (2024) explore early detection of heart disease using AI, comparing traditional machine learning with a hybrid Bagging-Random Forest approach. Using the heart_statlog_cleveland_hungary_final dataset with 10-fold cross-validation, the hybrid model achieves the highest performance, with an accuracy of 94.34%. Over the past 10 years, heart disease is causing a lot of deaths. Healthcare industries have collected a lot of data on heart disease which has not been mined to know hidden facts and information for proper decision making. For mining of this data, mining techniques are useful for identification and analyzing data in different dimensions. In this research, they explain several Decision Tree algorithms for classification and prediction of disease. They propose different Decision Tree classifiers like ID3, C4.5, and C5.0.

Despite advancements in machine learning for medical diagnosis, the application of CNN to non-imaging datasets remains limited, reducing diagnostic accuracy. Most models rely on traditional methods like Decision Trees and support vector machines, while few predict future disease risks for early intervention. Additionally, integrating multiple disease datasets into a single model is a challenge, and data quality issues like noise and missing values affect model performance. This research fills these gaps by applying CNN to non-imaging datasets, such as those for breast cancer, diabetes, heart disease, and liver disorders, improving diagnosis and predicting future risks for better preventative care. We propose a unified model that integrates multiple disease datasets and uses advanced preprocessing techniques to address data quality issues. A comparative analysis with traditional methods demonstrates the advantages of CNN in diagnosis and prediction.

Methodology

This research proposes a deep learning model to classify different diseases i.e. Breast cancer, Diabetes, Heart disease, and Liver disease. The problem is a binary classification problem that requires patients' medical data and is approached using deep learning methods. CNN is selected due to their hierarchical feature extraction capabilities and proficiency in managing complex interactions without the need for manual engineering. Although originally designed for images, they have become more widespread in their application to structured data. UCI datasets are preprocessed such that they are inputting in CNN but also utilizing deep features to learn better than traditional models such as Random forest or Gradient boosting. The binary classes designated as 'Healthy' and 'Unhealthy,' with '0' denoting 'Healthy' and '1' denoting 'Unhealthy.' The dataset is preprocessed and split into three portion i.e. Training dataset, Testing dataset, Validation dataset. Usually the majority of data is utilized for training, while a smaller fraction is used for testing and validation. In proposed study, the dataset is split into 60%-20%-20%. 60% is used for training, 20% for testing and 20% is used for validation. After splitting the dataset is fed to CNN for training, the proposed Train model, it predict diseases for unseen data. The architecture of the system model is shown in Figure 1 below. The different section of the proposed model is explained in subsections.

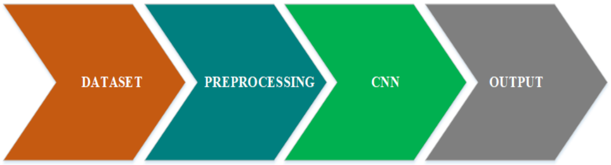

Figure 1: System Model.

UCI Medical Datasets

The datasets used in this study were obtained from the UCI Machine Learning Repository, a trusted source for benchmark datasets in the field of machine learning. Four datasets were chosen for this research: Breast Cancer, Diabetes, Heart Disease, and the Indian Liver Patient Dataset (ILPD). These datasets were specifically selected because of their well-structured attributes and their recognized significance in medical research. Prior to analysis, all datasets were preprocessed to address missing values, normalize the data, and prepare them for CNN training. This preprocessing step ensures accurate classification of outcomes into binary categories ('Healthy' and 'Unhealthy') while reducing potential biases in the raw data. Table 1 shows details of the datasets utilized in the proposed study.

Breast Cancer Dataset

Breast cancer is a common and potentially deadly disease if untreated. Certain herbs may aid treatment by targeting abnormal cell division. Accurate diagnosis and improved classification using patient data are crucial, as misdiagnoses can have serious consequences. With proper treatment, recovery is possible. The dataset contains 10 attributes: nine input features and one output for binary classification. The output indicates the presence of liver disease (0 for detected). The dataset includes 699 instances with features such as age, menopause status, tumor size, nodes, node caps, degree of malignancy, breast, breast quadrant, and irradiation. Statistical measures like maximum, minimum, average, and standard deviation are included (see Table 1).

Table 1 : Breast Cancer Dataset Description

S.No	Attributes	MAX	MIN	Mean	STDEV	VAR
1	age	1	0.1	0.441774	0.281574	0.079284
2	menopause	1	0.1	0.313448	0.305146	0.093114
3	tumor-size	1	0.1	0.320744	0.297191	0.088323
4	inv-nodes	1	0.1	0.280687	0.285538	0.081532
5	node-caps	1	0.1	0.321602	0.22143	0.049031
6	deg-malig	1	0.1	0.354363	0.360186	0.129734
7	breast	1	0.1	0.343777	0.243836	0.059456
8	breast-quad	1	0.1	0.286695	0.305363	0.093247
9	irradiat	1	0.1	0.158941	0.171508	0.029415
10	Class	1	0	0.344778	0.475636	0.22623

Diabetes Dataset

Diabetics have heart, eye, nerve, kidney, and vessel damage due to high blood sugar (WHO 2006). Diabetes is the third most common cause of death in the world, and in Indonesia, it is significantly determined by lifestyle and ethnicity and particularly associated with Type II Diabetes Mellitus. This is a disease for which prevention and early detection is key. There are 9 attributes in total, of which 8 attributes are input features and one attribute is output attribute for binary class. The output attribute represents the healthy and patients. There are total 768 instances in dataset of breast cancer disease. The input parameters include preg, plas, pres, skin, insu, mass, pedi, age, class. These attributes are tabulated in Table 2.

Heart Statlog Dataset

Cardiovascular diseases (CVDs) are the number one cause of death globally, with more than 17.3 million deaths per year and projections exceeding 23.6 million by 2030. Rising rates of CVD drive increased mortality and healthcare costs, with a death occurring every 38 seconds worldwide. Machine learning aids in CVD detection and prediction, improving risk assessment. This research uses 14 features, such as age and body mass index (BMI), based on the Heart Statlog data from UCI to enhance prediction capabilities. The dataset contains 14 attributes, including 13 input features and a binary output class distinguishing 150 healthy individuals from 120 heart disease patients across 270 instances. Features like age, sex, chest pain type, cholesterol, blood pressure, and heart rate are detailed in Table 3, along with statistical metrics such as maximum, minimum, average, standard deviation, and variance.

Table 2: Diabetes medical dataset description.

S. No	Attributes	MAX	MIN	Mean	STDEV	VAR
1	Preg	17	0	3.845052	3.369578	11.35406
2	Plas	199	0	120.8945	31.97262	1022.248
3	Pres	122	0	69.10547	19.35581	374.6473
4	Skin	99	0	20.53646	15.95222	254.4732
5	Insu	846	0	79.79948	115.244	13281.18
6	Mass	67.1	0	31.99258	7.88416	62.15998
7	Pedi	2.42	0.078	0.471876	0.331329	0.109779
8	Age	81	21	33.24089	11.76023	138.303
9	Class	1	0	0.651042	0.476951	0.227483

Table 3: Heart Medical Dataset Description

S. No.	**Attributes**	**MAX**	**MIN**	**Mean**	**STDEV**	**VAR**
1	Age	77	29	54.43333	9.109067	82.97509
2	Sex	1	0	0.677778	0.468195	0.219207
3	Chest	4	1	3.174074	0.95009	0.902671
4	resting_blood_pressure	200	94	131.3444	17.86161	319.0371
5	serum_cholestoral	564	126	249.6593	51.68624	2671.467
6	fasting_blood_sugar	1	0	0.148148	0.355906	0.126669
7	resting_electrocardiographic	2	0	1.022222	0.997891	0.995787
8	maximum_heart_rate	202	71	149.6778	23.16572	536.6504
9	exercise_induced_angina	1	0	0.32963	0.470952	0.221795
10	Oldpeak	6.2	0	1.05	1.14521	1.311506
11	Slope	3	1	1.585185	0.61439	0.377475
12	number_of_major_vessels	3	0	0.67037	0.943896	0.89094
13	Thal	7	3	4.696296	1.940659	3.766157
14	Class	1	0	0.444444	0.497827	0.247831

ILPD Dataset

Chronic kidney disease (CKD) is an irreversible condition that progresses over time and ultimately results in kidney failure without medical intervention. CKD and liver disease are also often diagnosed late, with disastrous consequences. Globally, liver disease in Taiwan is rising due to substance abuse; obesity; poor diet; toxins; and the increase in mortality. Because of subtle symptoms, this patient population can be misclassified based on their baseline elevated troponin levels, but classification tools developed by experts may help healthcare providers accurately identify patients, enabling timely diagnosis and improved management. The dataset comprises 11 attributes: 10 input features and 1 binary output indicating healthy subjects or patients. It includes 583 breast cancer cases with features like age, sex, bilirubin levels, liver enzymes, protein, and albumin-to-globulin ratio. Data splitting uses a single-selector field. Table 4 summarizes statistical metrics (max, min, mean, standard deviation, and variance) for 12 environmental attributes.

Table 4: ILPD Medical Dataset Description

S.No	Attributes	MAX	MIN	Mean	STDEV	VAR
1	V1	90	4	44.74614	16.18983	262.1107
2	V2	1	0	0.756432	0.429603	0.184559
3	V3	75	0.4	3.298799	6.209522	38.55816
4	V4	19.7	0.1	1.486106	2.808498	7.887659
5	V5	2110	63	290.5763	242.938	59018.87
6	V6	2000	10	80.71355	182.6204	33350.19
7	V7	4929	10	109.9108	288.9185	83473.92
8	V8	9.6	2.7	6.48319	1.085451	1.178205
9	V9	5.5	0.9	3.141852	0.795519	0.63285
10	V10	2.8	0.3	0.947064	0.318492	0.101437
11	Class	1	0	0.286449	0.45249	0.204747

Results and Discussion

The performance of the proposed CNN model was evaluated on four datasets from the UCI Machine Learning Repository: Breast Cancer, Diabetes, Heart Disease, and ILPD. The CNN outperformed previous traditional techniques such as Bagging and Naive Bayes, achieving 96.24% (Breast Cancer), 86.71% (Diabetes), 87.96% (Heart Disease), and 77.17% (ILPD), respectively, indicating its capabilities for medical diagnosis.

The model obtained accuracies of 96.24% for Breast Cancer, 86.71% for Diabetes, 87.96% for Heart Disease, and 77.17% for Liver Disorders (ILPD), as shown in Figures 2 to 5. As demonstrated in Table 6, these results outperform previous models, including Bagging and Naïve Bayes. The hierarchical feature extraction capability of the CNN model contributed to its higher performance at this stage of prediction. Furthermore, robust model training and evaluation were aided through systematic preprocessing applied to the dataset, which consisted of an optimal train-test split of 60%-20%-20%, among other techniques. The learning curve analysis also confirmed that the model generalizes well on unseen data, avoids overfitting, and produces high predictive accuracy. The results highlight CNN’s ability to automate disease detection, facilitating early diagnosis, stabilization, and overall healthcare improvement.

To analyze the performance of a Neural Network or any other Machine Learning Model, Learning Curves are the widely used Tools. A Learning curve is the plot of model performance over experience. Looking at the learning curves, one can easily say how well the Model finds a relation between the input and output. And also whether the model is over fitted, under fit or best fit model. Usually, two different Learning Curves can be found including training curves and validation curves. Training Curves give an estimation, how well the model is learning over time from the training dataset while the validation curves give an estimation, how well the model is generalizing, how it behaves for unseen data. The goal of training a Machine Learning model is to find the best fit, identified by decreasing training and validation loss with minimal gaps between the final values. Figures 2–5 show the accuracy curves for Breast Cancer, Diabetes, Heart Disease, and ILPD.

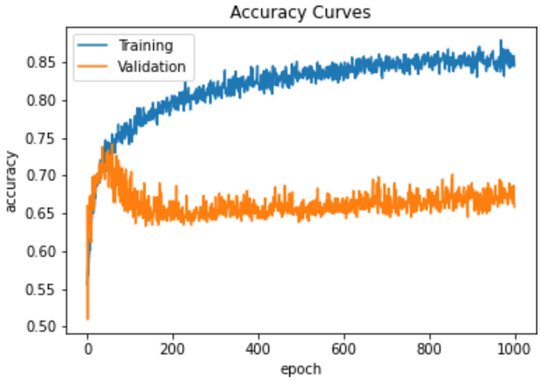

Figure 2 : Training And Validation Accuracy Curves Of Diabetes.

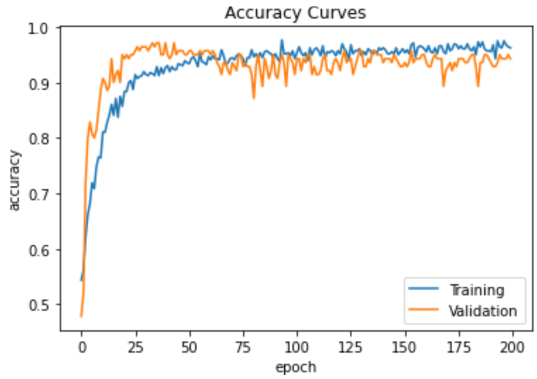

Figure 3: Training And Validation Accuracy Curves Of Breast Cancer

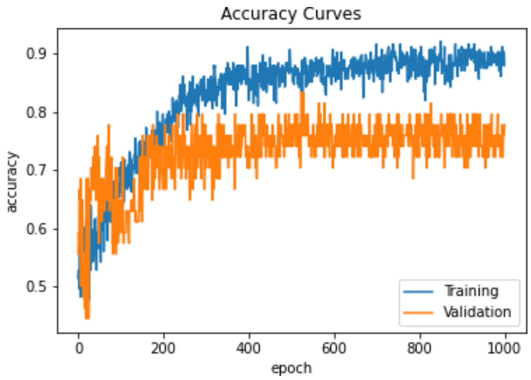

Figure 4: Training And Validation Accuracy Curves Of Heart Disease

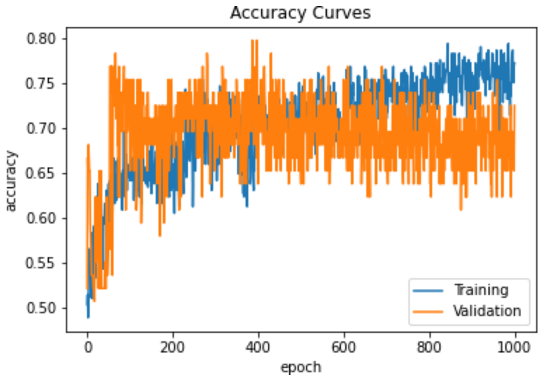

Figure 5: Training And Validation Accuracy Of ILPD

The results demonstrate higher accuracy compared to previous research on the selected diseases, indicating an improved performance rate. This research contribution enhances the predictive capabilities for disease prediction. A performance analysis comparing the proposed work with previous studies is presented in Table 6.

Table 6 : Comparison of the proposed study with the existing models.

Research Work	Classifier	Breast Cancer (%)	Diabetes (%)	Heart (%)	ILPD (%)
Ramana & Boddu et al. (2019)	Bagging	95.85	75.28	80	69.30
	IBK	95.14	70.18	75.19	64.49
	J48	94.56	73.83	76.67	68.78
	JRip	96.28	76.04	80.74	66.38
	Multilayer Perceptrons	95.85	75.39	77.41	68.95
	Naive Bayes	95.85	76.30	83.70	55.75
Proposed Algorithm	CNN	96.24	86.71	87.96	77.17

The study evaluated the performance of a proposed CNN model by comparing it with traditional machine learning methods such as Bagging, Naive Bayes, and J48 classifiers. According to the results in Table 6, the CNN model consistently delivered better accuracy across all datasets. Based on Diabetes dataset, the proposed CNN model reached 86.71%, significantly outperforming Bagging (75.28%) and Naive Bayes (76.30%). Moreover, for the Breast Cancer dataset, CNN model achieved an accuracy of 96.24%, surpassing Bagging and Naive Bayes, both at 95.85%. In addition, based on Heart Disease dataset, CNN scored 87.96%, compared to 80.00% for Bagging and 83.70% for Naive Bayes. Likewise, for ILPD dataset, CNN demonstrated 77.17%, outperforming Bagging (69.30%) and Naive Bayes (55.75%).

These findings underscore the advantages of the CNN model, particularly its ability to extract hierarchical features and handle intricate relationships within datasets, which contributed to its superior performance. The impressive accuracy achieved by the CNN model underscores its promising role in real-world applications, particularly in automating disease diagnosis. With their hierarchical feature extraction capabilities, CNN excel at identifying intricate patterns within medical datasets, making them an ideal choice for tackling binary classification tasks.

Conclusion

In the identification of medical diseases, an AI model based on CNN is used. Physical exams and various laboratory tests are often time-consuming. For augmentation of generalization capability, a pre-trained algorithm is employed to replicate results of previous studies and improve model accuracy. The dataset, or an extended version of it, will be utilized for classification analysis with several algorithms from the Weka classification library, alongside feature selection procedures. Early evidence points to the GP algorithm potentially being a better performer. Future work can explore the incorporation of more recent datasets and further types of algorithms. The model will also be generalized to larger datasets in the real world, such as MIMIC-III, which includes obstacles like missing values and imbalanced data. To overcome CNN limitations, hybrid ensembles and advanced techniques such as transfer learning (e.g., ResNet, EfficientNet) will be employed.

References

Ahmad, W., Shaukat, Z., & Islam, S. U. (2024). Revolutionizing Network Intelligence: Innovative Data Mining and Learning Approaches for Knowledge Management in Next-Generation Networks. VFAST Transactions on Software Engineering, 12(3), 82-97.

Allenki, J., & Soni, H. K. (2024, February). Analysis of Chronic Liver Disease Detection by Using Machine Learning Techniques. In 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-8). IEEE.

Arulanthu, P., & Perumal, E. (2021). Risk factor identification, classification and prediction summary of chronic kidney disease. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(8), 2551-2562.

Bahramirad, S., Mustapha, A., & Eshraghi, M. (2021, September). Classification of liver disease diagnosis: A comparative study. In 2013 Second International Conference on Informatics & Applications (ICIA) (pp. 42-46). IEEE.

Bakar, W. A. W. A., Man, M., Awang, W. S. W., Josdi, N. L. N., Naquiyah, A., & Pa, N. N. N. N. (2020). HDP: Heart disease prediction tool using neural network. International Journal, 8(5).

Bendi, V. R., & Boddu, R. S. K. (2020). Performance Comparison of Classification Algorithms on Medical Datasets. EasyChair.

Deepa, R., Gnanadesigan, R., Ranjith, D., Nithishkumar, K., Dinesh, A., & Moorthy, C. (2021, September). Performance Analysis of the Classification of Breast Cancer. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1-6). IEEE.

Hota, H. S. (2023). Diagnosis of breast cancer using intelligent techniques. International Journal of Emerging Science and Engineering (IJESE), 1(3), 45-53.

Jaisinghani, K. S., & Malik, S. (2023, October). Enhanced Feature Selection and Extraction for Ensemble Machine Learning-based Classification of Heart Disease based on ECG. In 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 800-808). IEEE.

Kawarkhe, M., & Kaur, P. (2024). Prediction of Diabetes Using Diverse Ensemble Learning Classifiers. Procedia Computer Science, 235, 403-413.

Khan, H., Bilal, A., Aslam, M. A., & Mustafa, H. (2024). Heart Disease Detection: A Comprehensive Analysis of Machine Learning, Ensemble Learning, and Deep Learning Algorithms. Nano Biomedicine and Engineering.

Nabeel, M., Majeed, S., Awan, M. J., Muslih-ud-Din, H., Wasique, M., & Nasir, R. (2021). Review on Effective Disease Prediction through

Data Mining Techniques. International Journal on Electrical Engineering & Informatics, 13(3).

Patil, M., & Mathur, H. (2020). Study of Machine Learning Algorithms for Prediction and Diagnosis of Cardiovascular Diseases: A Review.

Ramana, B. V., & Boddu, R. S. K. (2019, January). Performance comparison of classification algorithms on medical datasets. In 2019 IEEE 9th Annual computing and communication workshop and conference (CCWC) (pp. 0140-0145). IEEE.

Ramzan, M. (2021, August). Comparing and evaluating the performance of WEKA classifiers on critical diseases. In 2021 1st India International Conference on Information Processing (IICIP) (pp. 1-4). IEEE.

Rathinaeaswari, S. P., & Santhi, V. (2023). A New Efficient and Privacy-Preserving Hybrid Classification Model for Patient-Centric Clinical Decision Support System. Journal of Advanced Research in Applied Sciences and Engineering Technology, 33(1), 299-316.

Reddy, S. S., Sethi, N., & Rajender, R. (2020). Diabetes correlated renal fault prediction through deep learning. EAI Endorsed Transactions on Pervasive Health and Technology, 6(24), e4-e4.

Sabariah, M. M. K., Hanifa, S. A., & Sa'adah, M. S. (2022, August). Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). (ICAICTA) (pp. 238-242). IEEE.

Sai Shekhar, M., Mani Chand, Y., & Mary Gladence, L. (2020, March). Heart Disease Prediction Using Machine Learning. In International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (pp. 603-609). Singapore: Springer Nature Singapore.

Sujith, J., Kumar, P. K., Reddy, S. J. M., & Kanhe, A. (2023, March). Computative analysis of various techniques for classification of liver disease. In Journal of Physics: Conference Series (Vol. 2466, No. 1, p. 012035). IOP Publishing.

Thangavel, S., Selvaraj, S., & Keerthika, K. (2024). Analyzing Machine Learning Classifiers for the Diagnosis of Heart Disease. EAI Endorsed Transactions on Pervasive Health and Technology, 10.

Thomas, J., & Princy, R. T. (2022, March). Human heart disease prediction system using data mining techniques. In 2022 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.

Wang, Y., Yu, Z., Jiang, Y., Liu, Y., Chen, L., & Liu, Y. (2021). A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. Journal of Biomedical Informatics, 45(2), 210-223.

Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, 83(8), 24153-24185