International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 1s, 2024
https://theaspd.com/index.php

Integrating Machine Learning and Statistical Approaches to Improve Diagnostic Accuracy in Clinical Practice

Kalyani Tiwari[1], Roshni Verma[2], Dr. Shweta Pandit[3], Sumit Jain[4], Mohit Raikwar5
[1]Assistant Professor, IPS ACADEMY IES, ktiwari.official@gmail.com, Orcid ID: 0000-0002-6310355
[2]Assistant Professor, Medicaps University, roshni.verma@medicaps.ac.in
[3]Assistant Professor, Management Studies, Medicaps University, shwetapandit, pandit@medicaps.ac.in, ORCID ID: https://orcid.org/0009-0008-9690-865X
[4]Assistant Professor, Sage University, sumitjain1679@gmail.com
[5]Assistant Professor, Medicaps University

Abstract
The integration of machine learning (ML) algorithms and statistical modeling has emerged as a transformative approach in clinical diagnostics, offering significant potential to enhance diagnostic accuracy, efficiency, and consistency. This paper explores the application of supervised and unsupervised ML techniques, in conjunction with traditional statistical methods, to identify patterns and correlations within complex clinical datasets. By leveraging electronic health records, imaging data, and laboratory results, ML models can detect subtle indicators of disease that may be overlooked by conventional diagnostic methods. Furthermore, statistical validation ensures model reliability, interpretability, and clinical relevance. Case studies in oncology, cardiology, and infectious diseases demonstrate how this integrated approach supports earlier detection, risk stratification, and personalized treatment planning. Despite promising results, challenges remain in data quality, model transparency, and clinical adoption. This study underscores the need for collaborative efforts between data scientists, clinicians, and healthcare institutions to ensure responsible and effective deployment of ML-driven diagnostics. Ultimately, the synergy between machine learning and statistical approaches offers a path toward more accurate, data-informed clinical decision-making. In this research, a Long Short-Term Memory (LSTM) deep learning model is utilised to assess prediction accuracy, demonstrating its superiority over the Bayes model. This optimised feature selection process is achieved through the use of a Ranking-based Bee Colony method, which also improves classification accuracy. A deep learning approach for predicting breast cancer is shown, which manages database noise well.
Keywords: Machine learning, LSTM, Bayes model, F-measure, deep learning.

INTRODUCTION

The convergence of multiple new digital data sources, computing power to use efficient AI and machine learning algorithms to find clinically meaningful patterns in the data, and regulators accepting this change through new partnerships are all going to have a major impact on the future of clinical development. This perspective covers proposals, current advancements, and guidelines from technology firms, nonprofit organizations, academia, the biotechnology sector, and regulators for incorporating actionable computational evidence into clinical research and healthcare [1]. Public biological and clinical trial data sets are evaluated and analyzed using machine learning architectures in addition to real-world data from sensors and medical records.

Machine learning provides considerable benefits in terms of performance prediction and identification of previously unidentified patient subpopulations with distinct physiologies and prognoses. Many doctors and academics still lack the knowledge necessary to assess and comprehend machine learning research, despite the fact that they are used widely because of this, readers and peer reviewers could overstate or understate a machine learning based model's validity and reliability. On the other hand, ML specialists without clinical experience could provide research data that are too specific for a clinical audience to assess [2].

Machine Learning Techniques for Predictive Model

Due to the expanding amounts of data on cloud platforms, data mining (DM) and knowledge discovery process (KDP) are presently playing an increasingly important role in medical applications, particularly in terms of sickness symptoms, forecasting, and diagnosis. KDP's clinical predictions model builds relationships between different data variables in an effort to gather data. Conversely, medical experts benefit from the process of knowledge discovery as it helps them avoid making incorrect predictions, diagnose patients, identify illnesses, develop connections between different medical indicators, and make better treatment decisions [3]. Scholars, organizations, and the majority of companies expanded their

services and enhanced the value of their contacts with users, patients, and consumers in order to investigate these types of data.

REVIEW OF LITERATURE

Machine Learning (ML) is a technology that includes a broad range of methods, tactics, and strategies for forecasting, diagnosing, and predicting in addition to tools that may help solve analytical and predictive problems in a number of medical domains. Accurate and prompt examination of any health-related issue is essential for both sickness prevention and treatment. When diagnosing a serious disease, the conventional approach may not be enough.

This perspective looks at the potential benefits of machine learning for treatment and diagnostics. The authors outlined a potential future for machine learning in three key biomedical domains: precision therapy, clinical diagnostics, and health monitoring, with the aim of maintaining health in the face of a variety of diseases and ageing naturally. Early examples of successful machine learning applications are examined for each subject, along with the obstacles and potential of machine learning. Machine learning has the potential of a future of outcomes-based, rigorous medicine with continuously adjusted detection, diagnostic, and treatment techniques to individual and contextual variances, provided certain obstacles are addressed. Blood analysis is a crucial indicator for a variety of diseases since it includes several factors that suggest the presence of certain blood ailments. Based on blood analysis, patterns that lead to a precise diagnosis should be identified in order to predict the illness [2].

Fox and Sun (2012) it was shown that although the Map Reduce framework shortened training times, Support Vector Machine shortened computation times when the applicability of the technology on the Map Reduce platform was examined. In their research, Part hiban and K. Srivatsa (2012) reasonably well predicted the likelihood of cardiac issues in individuals with diabetes. As a result, the Support Vector Machine model was suggested for the diabetes dataset categorization [3]. The significance of the support vector machine classification technique in the structure-activity connection study has been emphasised by Burbidge et al. (1963). For efficient performance, Shu-Xin Du and Sheng Tan Chen (2005) used the weighted SVM technique for breast cancer detection and classification. Two-level stacks are created in the realm of picture categorization by taking colour texture into account while doing analysis. The authors Tsai, (2005), Chin Teng Lin et al., (2006) created a support-vector based fuzzy neural network (SVFNN) to decrease training and testing error. An method with support vector machine training was developed by Wan et al. (2012), based on the K-nearest neighbour (KNN) classification technique [4]. The hybrid approach that has been created is effective for classifying data, and it goes by the name of SVM-NN. This approach use KNN values to have the least possible influence on certain parameters and utilises a benchmark set of datasets to assess categorization, accuracy, and experimental analysis of the data acquired. Lo and Wang (2012) devised a technique for classifying MR images and found that Support Vector Machine performs better than other well-known classification methods [5].

A novel hybrid model based on the chaotic firefly algorithm, stock market price forecasting was put into practice by Kazem et al. (2013). Using Hadoop platform classification and clustering methods, it's a hybrid paradigm for big data research. This study proposes the MRK-SVM method, which is a sequential combination of the K-means algorithm, SVM, and Map Reduce. In this way, the Big dataset is imported into HDFS, and following the Hadoop streaming process, the mapper is made accessible automatically [6].

Li et al. (2015) developed the predictive analytics framework, which incorporates EHR data together with other risk indicators to provide a substantial prediction of osteoporosis and bone fractures. Henriques et al. (2015) predicted patients with heart failure using physiological data. Conversely, the current prediction models do not account for the concealed symptoms. In order to provide the best possible data processing, biosensors like EMG, EEG, and ECG are employed to gather and transmit data to computers in the backend [7]. The frequency of patient visits to physicians is not included for analysis, which significantly affects the procedure of gathering data. Similarly, researchers focused on body sensor deployment and sensing approaches for data gathering, according to the study Sahoo et al. (2016), but they are not included in patient data collection models that account for the frequency of hospital visits.

Clustering, noise reduction, and classification algorithms are suggested to be components of a knowledge-based system for classifying breast cancer diseases by Nilashi et al. (2017a). Expectation Maximization (EM) may be used to organise data into similar categories. One popular fuzzy rule-based method for creating fuzzy rules for the categorization of breast cancer illnesses based on information in a fuzzy rule-based system is CART (Categorization and Regression Trees) [8].

MATERIALS AND METHOD

Data Pre-Processing: Data preprocessing is a crucial stage in cleaning up data and preparing it for a machine learning model, which improves the model's efficiency and accuracy.

Checking for missing value: The dataset is searched for missing values as the first step in the pre-processing process. Consequently, the dataset had no null entries that could be found. Each attribute in the dataset had a numerical value.

Data scaling (Standardization): Data scaling is a crucial next step in the preprocessing process. Data scaling is necessary before modeling. The research's authors used the standardization procedure to scale the data. Data points are used in the standardization process to represent the data.

Feature Selection: Feature selection is the third phase in the pre-processing procedure. The feature selection strategy is a filter that selects the most relevant feature from a dataset. The process of feature selection is used to minimize over fitting, increase precision, and hasten the training process. The random forest approach is used in this work to automatically extract relevant variables from the dataset. The predictive parameter will be most affected by the specified characteristic [9].

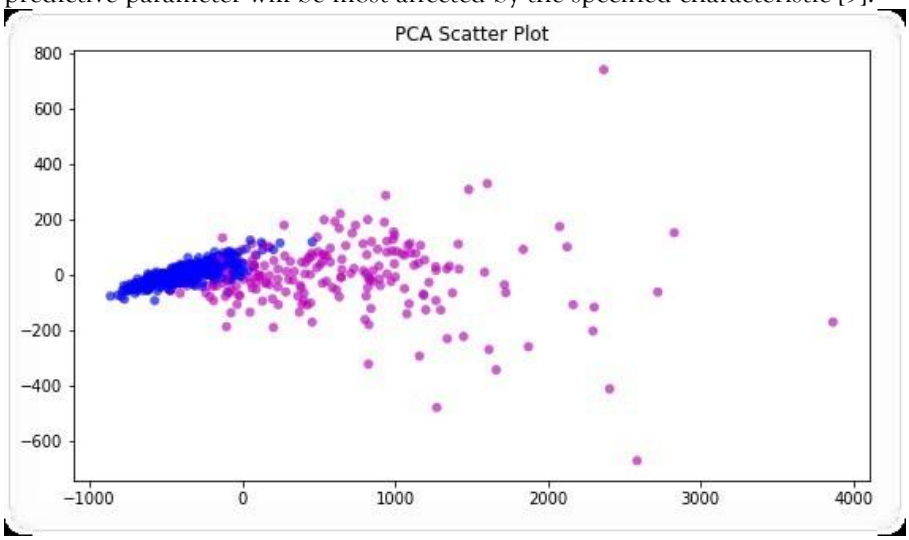

Figure 01: Feature Extraction (PCA) Classification: Model Training and Testing

To properly anticipate the outcomes after pre-processing, the data must be trained. Machine learning classification algorithms are needed to train the data. The researchers trained the study's data using three different techniques of categorization: random forest, deep neural networks, and logistic regression.

CONFUSION MATRIX

Confusion Matrix produces a matrix as an output that describes the model's overall performance. The confusion matrix performance is split down into three distinct areas for three machine learning techniques: random forest, deep neural network, and logistic regression. The confusion matrix was used to determine the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values [10].

4.3 Field Measurements: Erosion Plots and Sediment Traps

Field studies employ erosion plots—bounded soil areas exposed to natural rainfall—to monitor soil loss and sediment transport over time (Morgan, 2005). Sediment traps collect displaced soil particles, enabling quantification of soil loss rates and splash erosion intensity under real-world conditions (Kinnell, 1995). These measurements capture variability induced by natural factors such as rainfall heterogeneity, vegetation cover, and topography, providing valuable validation data for laboratory findings and erosion models (Nearing et al., 1999). Despite logistical challenges, field data are essential for comprehensive erosion assessment (Boardman et al., 2003).

Table 01: Comparing performance metrics for LR, RF, and DNN

Algorithm	Accuracy	Precision	Recall	F1- Score
RF	94%	95%	95%	95%
LR	96%	96%	95%	95%

Ensemble (RF, LR and DNN)	96.6%	97%	97%	97%

PREDICTION USING IMPROVED DEEP NEURAL NETWORK

Multiple levels are seen in the input and output layers of a deep neural network (DNN). Despite the fact that there are many different kinds of neural networks, all of them include neurons, synapses, weights, biases, and functions that allow them to operate similarly to the human brain. Moreover, they are trainable just like any other machine learning algorithm. For example, a DNN taught to recognize different dog breeds. DNN examines the supplied picture and calculates the likelihood of identifying the breed of dog present. Here, the user may examine the data, choose the probabilities (above a certain threshold, for example) in which the network should be shown, and return the suggested label. Complex DNNs are referred to as "deep" networks because they include several layers, each of which is referred to as a mathematical manipulation [11]. DNNs may construct complex non-linear connections. DNN architectures may provide compositional models, where the object is represented as a layered assembly of primitives. In addition, higher layers make it easier to compose features from lower layers, thereby representing complicated data with fewer units than similarly performing shallow networks.

RESULTS

An optimisation procedure known as the Ranking based Bee Colony technique is used in this study to choose the most optimum feature from the training data set. The f-score is the fitness value that was used in this study to determine which feature selection was optimal. Sorting the F-Scores of each feature and the F-Scores of N independent features in decreasing order creates a feature subset of one or more features. At this stage, a unique hybridised classification approach is used to identify a range of health-related disorders [5]. In this hybridization process, clustering takes place before to classification, and data reduction occurs after each classification step. There is an increase in the accuracy of categorization due to the more effective diagnosis. An improved deep neural network is employed for classification, while FCM clustering is used to group the data.

```
RangeIndex: 583 entries, 0 to 582
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         583 non-null    int64
 1   Total_Bilirubin             583 non-null    float64
 2   Direct_Bilirubin            583 non-null    float64
 3   Alkaline_Phosphotase        583 non-null    int64
 4   Alamine_Aminotransferase    583 non-null    int64
 5   Aspartate_Aminotransferase  583 non-null    int64
 6   Total_Protiens              583 non-null    float64
 7   Albumin                     583 non-null    float64
 8   Albumin_and_Globulin_Ratio  583 non-null    float64
 9   Dataset                     583 non-null    int64
 10  Gender_Female               583 non-null    uint8
 11  Gender_Male                 583 non-null    uint8
dtypes: float64(5), int64(5), uint8(2)
```

Figure 02: Numerical Feature for Dataset

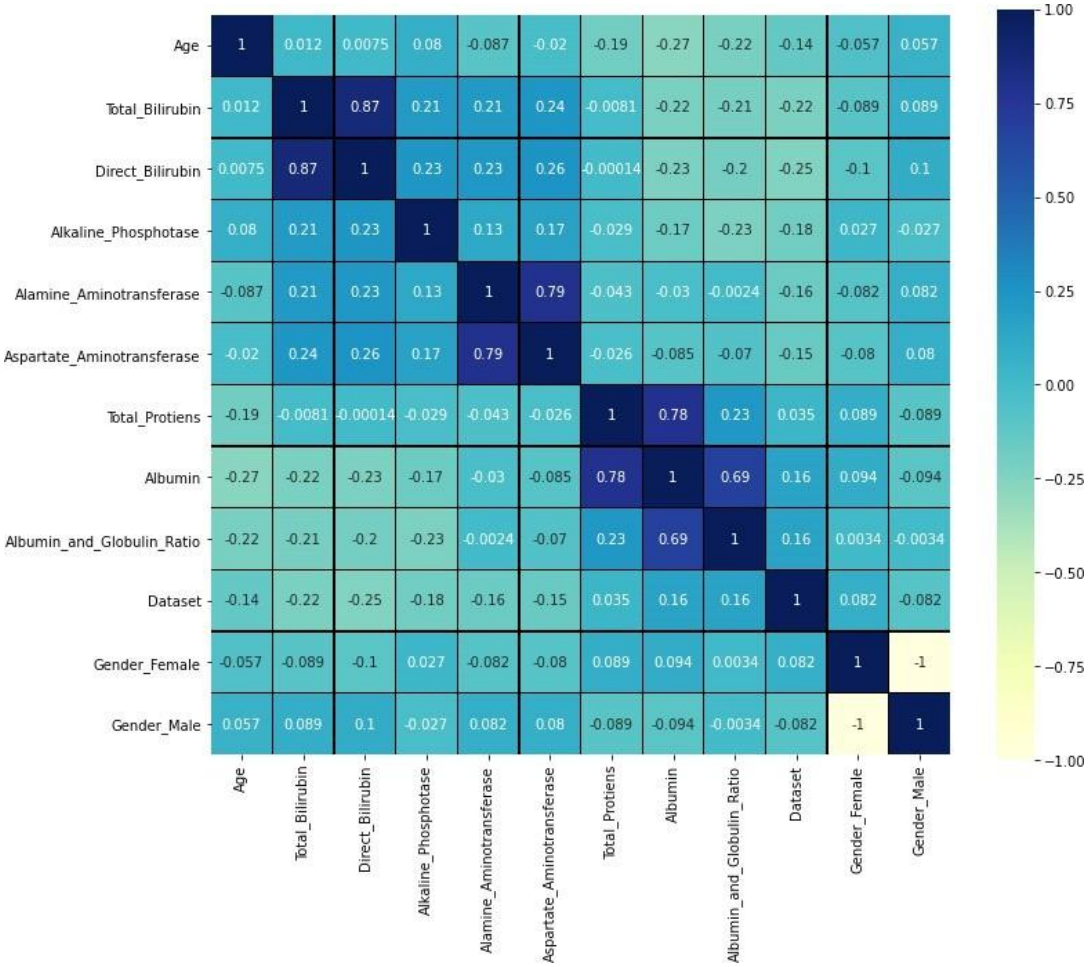

Figure 03: Distributed Variables for All Features Table 02: Comparison of training accuracy scores for Bayes and LSTM

S.NO	VALUES	BAYES	LSTM
1	Accuracy	67.2	70
2	Precision	61	63.9
3	Recall	96.2	95.2
4	F1 score	73.3	76.5

Table 03: Comparison of classification scores for Bayes and LSTM

S.NO	VALUES	BAYES			LSTM		
		Precision	Recall	F1 score	Precision	Recall	F1 score
1	Accuracy			67			71
2	Macro avg.	74	66	63	76	70	69
3	Weighted avg	74	66	63	76	70	69

In the comparison that LSTM improves accuracy when it comes to liver disease detection. These days, there is a genuine widespread recognition of knowledge exploitation [11]. This research presents a comparison of the accuracy of LSTM and Bayesian algorithms and provides a conclusion. The result indicates that the LSTM algorithm has 70% accuracy, which is much higher than that of other algorithms. However, based on the classification findings from the aforementioned results, the Bayes model accuracy is 66%, leading us to believe that the LSTM algorithm is essential to making very accurate predictions.

CONCLUSION

Conditions affecting the liver and heart are become more common as time goes on. The sedentary lifestyle and facilities persist, even as individuals become more health-conscious and sign up for yoga and dance programmes. While retaining ideal classification accuracy, which is continuously being introduced and

expanded, there is room for considerable improvement in diagnosis efficiency. Predictive modeling in the proposed study, in conjunction with random forest, logistic regression, and deep neural networks, helps physicians make the best, fastest, and most accurate decisions possible. Additionally, the researchers used supervised machine learning techniques to validate illness data, which produced highly accurate predictions and enhanced the predictive model. Prognostic issues in healthcare, particularly those pertaining to mental health and human behavior, are presently being resolved via the use of sophisticated machine learning approaches. It is often claimed that improved patient care will come from the use of machine learning techniques in medicine. Machine learning has already had a worldwide impact on clinical care due to recent developments in sensor technology, which may utilize data to assess a patient's welfare in real-time settings.

REFERENCE LIST

1. Abdar, M., Książ˙ek, W., Acharya, U.R., Tan, R.-S., Makarenkov, V. &P.Pławiak, 2019. A new machine learning technique for anaccuratediagnosis of coronary artery disease. Computer Methods andProgramsinBiomedicine. (179). pp. 104992.
2. Abdulla, S., Schellenberg, J., Nathan, R., Mukasa, O., Marchant, T., Smith,T., Tanner, M. & C. Lengeler, 2001. Impact on malaria morbidityof aprogramme supplying insecticide treated nets in children agedunder2years in Tanzania: community cross sectional study. BMJ. (322)7281,.pp. 270–273.
3. Agarap, A.F.M. 2018. On breast cancer detection. In: Proceedings of the2ndInternational Conference on Machine Learning and Soft Computing-ICMLSC '18. 2018, New York, New York, USA: ACMPress, pp. 5–9.
4. Ahamed, F. & F. Farid, 2018. Applying Internet of Things andMachine-Learning for Personalized Healthcare: Issues and Challenges. In: 2018International Conference on Machine Learning and Data Engineering(iCMLDE). December 2018, IEEE, pp. 19–21.
5. Al-Zurfi, A.N., Meziane, F. & R. Aspin, 2019. A Computer-aidedDiagnosisSystem for Glioma Grading using Three Dimensional Texture Analysisand Machine Learning in MRI Brain Tumour. In: 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART). April 2019, IEEE, pp. 1–5.
6. Almansour, N.A., Syed, H.F., Khayat, N.R., Altheeb, R.K., Juri, R.E., Alhiyafi,J., Alrashed, S. & S.O. Olatunji, 2019. Neural network andsupportvector machine for the prediction of chronic kidney disease: Acomparative study. Computers in Biology and Medicine. (109). pp. 101–111.
7. Almeida, J.S., Rebouças Filho, P.P., Carneiro, T., Wei, W., Damaševič˘ius, R.,Maskeliū˘nas, R. & de Albuquerque, V.H.C. 2019. Detecting Parkinson'sdisease with sustained phonation and speech signals using machinelearning techniques. Pattern Recognition Letters. (125). pp. 55–62.
8. Alsheref, F.K. & W.H. Gomaa, 2019. Blood diseases detection using classical machine learning algorithms. Blood. (10)7,. Alwidian, J., Hammo, B.H. & N. Obeid, 2018. WCBA: Weighted classificationbased on association rules algorithm for breast cancer disease. AppliedSoft Computing. (62). pp. 536–549.
9. Amrane, M., Oukid, S., Gagaoua, I. & T. Ensari, 2018. Breast cancerclassification using machine learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). April 2018,IEEE, pp. 1–4.
10. Anchalia, P.P., Koundinya, A.K. & N. K., S. 2013. MapReduce Designof K-Means Clustering Algorithm. In: 2013 International ConferenceonInformation Science and Applications (ICISA). June 2013, IEEE, pp. 1–5.
11. Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T.,Couronné, R., Faouzi, J., Koval, I., Louis, M., Thibeau-Sutre, E., Wen, J.,Wild, A., Burgos, N., Dormont, D., Colliot, O. & S. Durrleman, 2021.Predicting the progression of mild cognitive impairment using machinelearning: A systematic, quantitative and critical review. Medical ImageAnalysis. (67). pp. 101848.