

A Comparative performance of Machine Learning models with cross validation techniques for the prediction lever disease.

Madhab Paul Choudhury
Research Scholar, PhD(engg.)
IIT ISM Dhanbad
Jharkhand

J Paul Choudhury
Retired Professor(IT), Kalyani Govt Engg. College
Kalyani,, West Bengal
Professor, Narula Institute of Technology, Kolkata

Abstract.

Artificial Intelligence (AI) has changed something so that it is much better in various aspects of our lives, offering solutions to numerous problems and bridging gaps between reality and business. Within the domain of AI, emerging technologies such as machine learning and deep learning models have played an important role in transforming the way so that one can analyze data, make decisions, and can take action or give attention to difficult situations or problems, with an objective to understand them to find solutions. With the exponential growth of data usage and storage, these technologies have assumed a vital role in data analytics, storage management, and decision-making processes. Through digital transformation, reshape of industries becomes possible, worldwide services in all areas including health care can be modified for the betterment of human life.

Liver disease encompasses a wide range of conditions that impair the liver's ability to function properly, leading to serious health issues. Risk factors for liver disease include viral infections (e.g., hepatitis B and C), excessive alcohol consumption, obesity, metabolic disorders, and genetic predisposition. Symptoms vary depending on the disease's severity but commonly include fatigue, jaundice, abdominal pain, and swelling. In the early stages, liver disease is often without symptoms, making early detection and intervention crucial. Advanced liver disease can result in liver failure, and in some cases, a liver transplant may be necessary. If the lever disease can be detected in early life, certain preventive measures can be taken so that the patient can recover from lever disease and can enjoy a real and happy life.

This paper intends to create a lever disease predictive model utilizing medical data and investigating the impact of various factors on the condition using machine learning techniques and cross validation techniques. Millions of individuals throughout the world suffer from the common chronic illness known as lever disease. The results of patients and public health initiatives can be considerably improved by early detection and an understanding of the underlying causes. To do this, a large dataset of medical records from people with and without lever disease that included a variety of demographic, lifestyle, and clinical factors have been taken. To pre-process the data and identify useful features, feature engineering approaches have been used. Accurate prediction models for lever disease risk assessment have been created using a variety of machine learning methods, such as Decision Trees, Random Forests, and Support Vector Machines. The main causes of the development of lever diseases have been determined by looking into disease influence measures.

Here machine learning models and cross validation techniques of machine learning models have been applied for selecting a proper model for prediction of lever disease of the person. Under machine learning models Random forest, decision tree, gradient boosting, KNN(K nearest neighbor), logistic regression have been used. Under cross validation techniques, k fold cross validation, cross validation using estimator values, stratified k fold cross validation, leave one out cross validation, hold out method and recursive feature elimination method have been

used.

Keywords.

machine learning models, Random forest, decision tree, gradient boosting, KNN(K nearest neighbor), Cross validation techniques, k fold cross validation, cross validation using estimator values, stratified k fold cross validation, leave one out cross validation, hold out method and recursive feature elimination method.

1.Introduction.

Early detection of liver disease is essential for effective treatment. Advanced imaging methods, such as elastography (including FibroScan), aid in assessing liver stiffness, which may indicate fibrosis or cirrhosis. Blood biomarkers like ALT, AST, and specialized liver function tests are commonly used to evaluate liver health. Recently, researchers have explored non invasive biomarkers and genetic markers for their potential in predicting disease progression, especially in conditions like NAFLD and nonalcoholic steatohepatitis (NASH).

2. Literature Review.

The research paper[1] focuses on utilizing emerging technologies to predict the progression of Periodontitis, a prevalent oral disease. Machine learning classifiers are applied to analyze and forecast the disease's development. The study also incorporates techniques like cross-validation, feature extraction, and ensemble learning, which are evaluated for their effectiveness. The performance of various classifiers, including Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, K Nearest Neighbours (KNN), and Decision Tree, is compared. These classifiers are tested on a dataset of 1000 periodontitis patients, yielding remarkable accuracy rates: 95.5%, 100%, 100%, 100%, 99.5%, and 99% for chronic localized and chronic generalized periodontitis classification, respectively.

Early detection and understanding the underlying causes of diabetes can significantly enhance patient outcomes and public health initiatives [2].To achieve this, a comprehensive dataset comprising medical records from individuals both with and without diabetes is collected, encompassing various demographic, lifestyle, and clinical factors. Feature engineering techniques are applied to pre process the data and identify relevant features. Accurate prediction models for assessing diabetes risk are developed using diverse machine learning algorithms, including Decision Trees, Random Forests, and Support Vector Machines. Additionally, by examining disease influence measures, the primary contributors to diabetes development are identified.

The authors [3] propose a framework for multi-class skin disease classification that combines a robust convolution neural network (CNN), MobileNet V2 (MNV2), and Long Short Term Memory (LSTM) to enhance the accuracy and reliability of diagnoses. This approach leverages LSTM's ability to manage multi-class classification tasks and CNN's strength in automatically learning distinguishing features from raw skin images. The proposed method surpasses previous models, achieving an accuracy of over 87%, using the HAM10000 dataset for evaluation.

Automated systems in healthcare now generate clinical records, predict infections, assist in diagnostics, and enable continuous patient monitoring [4].The integration of Internet of Things (IoT) devices and Machine Learning (ML) algorithms has been pivotal in this transformation. However, the performance of AI algorithms can vary across different datasets, leading to inconsistencies in predictive outcomes. Such variability poses potential risks to clinical decision-making processes. This paper presents key features of prominent ML algorithms used for classification and prediction, illustrating their applications in the healthcare sector.

Machine learning models can analyse gene expression data from DNA microarrays to predict the presence of lung cancer in patients [5]. Collective Random Forest and Adaptive Boosting are utilized to identify the contributing factors. Kernel Principal Component Analysis (KPCA) is employed for feature reduction, followed by calculating the correlation between each feature

and the target using KPCA's statistical parameters. One method of assessing a classification model's accuracy is by determining the proportion of correct predictions for a given dataset. In this study, the proposed technique is validated using a dataset containing lung cancer-related information, specifically GSE4115 from the Gene Expression Omnibus (GEO) database, along with its expression profiles. The results highlight the effectiveness of the Identification of Lung Cancer (IOLC) model, which achieves an accuracy of 81%, a precision of 81.2%, a recall of 78.9%, an F-Measure of 77.7%, and an error rate of 0.29%.

The advancement of intelligent and sustainable healthcare services is significantly propelled by the integration of Internet of Things (IoT) devices and cloud infrastructure [6]. By adopting a cloud-centric IoT paradigm, machine learning methods can enhance the capabilities of smart healthcare systems. In this approach, data collected from IoT devices are securely transmitted to the cloud for further analysis and processing. Machine learning algorithms then analyse this data to extract valuable insights, facilitating more informed healthcare services. Continuous training and updating of these models with the vast amounts of data stored in the cloud enable progressive improvements in their accuracy and performance. This framework aims to achieve intelligent and sustainable healthcare solutions by effectively combining machine learning techniques with cloud-centric IoT.

Machine learning has become an essential tool in the medical field, utilizing data patterns to enhance diagnostic accuracy. This paper [7] aims to minimize misdiagnosis and improve patient outcomes by developing a predictive model for cardiovascular diseases using machine learning techniques. The study focuses on accurately classifying cardiovascular conditions, assisting healthcare professionals in making informed decisions for timely and targeted treatments. The findings highlight the significant role of machine learning in medicine, demonstrating its potential to lower mortality rates associated with cardiovascular diseases. By applying ensemble classification methods to a dataset of heart disease patients, the research seeks to enhance prediction accuracy.

This study explores the application of Pathology Learning and Modeling (PaLM) methodologies to advance biomedical machine learning research for interpreting medical data [8]. PaLM involves developing and applying machine learning algorithms to analyze pathology data, including histological images, molecular pathology information, and clinical pathology details. By harnessing machine learning capabilities, PaLM enables researchers to uncover hidden patterns, correlations, and insights within complex medical datasets. The authors aim to enhance disease diagnosis, prognosis, and treatment planning through precise and efficient analysis of pathology data using machine learning models and algorithms. Integrating molecular pathology data, clinical data, and pathology images provides a comprehensive understanding of diseases and supports personalized patient management. PaLM approaches facilitate automated image analysis and segmentation, extraction of relevant features, and identification of disease-specific patterns.

This study presents an advanced approach for kidney stone prediction using a deep learning framework that integrates the YOLO v7 model with the Energy Valley optimizer for image segmentation and the Pulse Couple Neural Network (PCNN) for classification [9]. Initially, the YOLO v7 model is employed to detect and localize kidney stones for segmentation. To enhance its performance, hyper parameter tuning is conducted, allowing the model to effectively learn and generalize from data. The Energy Valley optimizer, inspired by energy valleys in physics, is introduced to further optimize YOLO v7, improving its segmentation accuracy. Additionally, the PCNN model is utilized as a classification framework, leveraging pulse-coupled oscillators to analyse the segmented kidney stone regions and identify their characteristics. The proposed methodology demonstrates exceptional performance across multiple evaluation metrics, achieving a precision of 98.58%, recall of 99.17%, accuracy of 98.88%, F1-score of 97.42%, and specificity of 98.23%.

This study evaluates the performance of various machine learning and deep learning models—including convolution neural networks (CNNs), decision trees, random forests, extra trees classifiers, dense models, and hybrid CNN-LSTM architectures—in classifying electrocardiogram (ECG) signals for disease detection [10]. An extensive review of existing literature on machine learning applications in ECG signal processing and healthcare is conducted. The research utilizes a meticulously curated and annotated dataset comprising ECG signals from both healthy individuals and patients with significant disorders, ensuring a comprehensive representation of the target population. In binary classification tasks, CNN and CNN-LSTM models consistently outperform other algorithms, achieving high accuracy, F1-scores, and AUC-ROC values. These models effectively distinguish between disease and non-disease categories, demonstrating their potential utility in clinical practice. The findings underscore the efficacy of CNN and CNN-LSTM models in enhancing diagnostic accuracy, thereby contributing to improved patient outcomes. The study also offers recommendations for further research and development in ECG signal processing, addressing challenges and considerations pertinent to the implementation of these algorithms.

The authors have developed a risk assessment model that utilizes convolution neural networks (CNNs) to evaluate a patient's likelihood of hospitalization or mortality due to heart failure within a large health maintenance organization [11]. This study highlights the potential of deep learning algorithms for the early detection of heart disease. The primary goal of the research is to assess the accuracy of diagnosing cardiac conditions in individuals. Instead of determining the optimal partition order, the recursive partitioning process follows a greedy approach for reordering partitions. The proposed system leverages a CNN model to process input datasets for disease prediction, incorporating pre processing, feature extraction, and classification techniques to analyze data and generate meaningful insights. By applying dimensionality reduction, the model enhances prediction accuracy using the same dataset. Notably, many conventional AI-based classification algorithms fail to surpass the accuracy levels achieved by Lasso or Ridge regression, both of which consistently deliver superior results.

The proposed approach by the authors integrates Long Short-Term Memory (LSTM) for sequence learning and Convolution Neural Networks (CNNs) for extracting non-linear features [12]. This methodology enhances the analysis of the Mammographic Image Analysis Society (MIAS) dataset, achieving improvements of 5% in accuracy, 6% in precision, and 4.6% in recall. The effectiveness of this approach is further validated using the INBREAST and Digital Database for Screening Mammography (DDSM) datasets through various performance metrics. Compared to existing methods, the proposed approach demonstrates a 2-3% increase in accuracy, a 2% improvement in precision, and a 3-4% boost in recall for the DDSM dataset. Similarly, in the INBREAST dataset, it enhances accuracy by 3-4%, precision by 2-3%, and recall by 4%, showcasing its effectiveness in mammographic image analysis.

Mammography has demonstrated promising outcomes with the application of deep learning technologies in the quantitative evaluation of parenchymal density, categorization, detection, diagnosis, and breast cancer risk prognosis, enabling more precise patient management [13]. Additionally, deep learning has streamlined the interpretation process, reducing both interpretation time and workload. However, more comprehensive research is needed to definitively confirm the effectiveness of deep learning. This article examines the classification of mammograms using deep learning techniques and explores their potential for mammography interpretation, along with the challenges they face in real-world applications. The proposed method combines Long Short-Term Memory (LSTM) for sequence learning and Convolution Neural Networks (CNN) for nonlinear feature mapping to enhance the accuracy of the Mammographic Image Analysis Society (MIAS) dataset by 5%, precision by 6%, and recall by 4.6%. Experiments using the INBREAST and Digital Database for Screening Mammography (DDSM) datasets validate the proposed approach, demonstrating

improvements in comparison to existing methods. Specifically, it improves accuracy by 2-3%, precision by 2%, and recall by 3-4% on the DDSM dataset, and boosts accuracy by 3-4%, precision by 2-3%, and recall by 4% on the INBREAST dataset.

This paper presents a comparative analysis of segmentation and feature extraction methods for detecting lung cancer [14]. Various segmentation techniques, such as Thresholding, global Thresholding, and watershed segmentation, are implemented and assessed. Additionally, feature extraction is applied to improve the performance of these segmentation techniques. The proposed method is compared with existing approaches to showcase its effectiveness and potential for better lung tumour detection. The study uses five images for analysis, and the results indicate that the proposed segmentation and feature extraction techniques can achieve higher accuracy, potentially aiding in the early detection of lung cancer.

Segmentation, detection, and classification are essential stages in digital imaging pathology labs for analyzing MRI brain tumour regions. This study focuses on medical image analysis and classification using a convolution + ReLU algorithm, which integrates convolutional techniques with ReLU optimization [15]. The research employs a robust and efficient convolution + ReLU approach on the BraTS 2020 dataset, significantly reducing segmentation time compared to other optimization methods. Furthermore, it demonstrates outstanding performance metrics, achieving 99.8% precision, 99% recall, and a 99.3% F-measure. Convolution neural networks (CNNs) utilizing the convolution + ReLU activation function enhance learning speed and improve tumour analysis performance. In the experimental phase, the implemented convolution + ReLU model achieved an impressive accuracy of 99.8%, surpassing existing methodologies.

Researchers process extensive and complex healthcare data using various deep learning techniques, enabling medical professionals to predict diseases effectively [16]. This study introduces a model designed to identify cardiac disorders, aiming to benefit numerous individuals globally. The proposed model enhances the Convolution Neural Network (CNN), referred to as Custom CNN (C-CNN), and demonstrates superior performance compared to previously published methods.

The authors have concentrated on machine learning (ML) algorithms for cancer prediction, which is impacted by various performance metrics [17]. By utilizing widely used ML techniques such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Linear Regression, Decision Tree, and Naive Bayes, the study evaluates the accuracy of cancer prediction. Their research provides insights and recommendations on applying machine learning techniques in clinical environments to enhance cancer detection and patient care.

In this study, the authors propose enhancing a MobileNet base model by fine-tuning it with additional features to improve brain tumour detection [18]. The model's precision and accuracy are increased by restructuring its layers. Pre-processing techniques are applied to MRI images to enhance their quality, and data augmentation is employed to expand the dataset size, thereby improving the model's training process. The results indicate that the proposed model outperforms other convolution neural network (CNN) models, including VGG16, Xception, and ResNet50, demonstrating the potential of deep learning in detecting brain tumours.

Since insulin plays a crucial role in regulating various properties of plasma, including water, enzymes, proteins, vitamins, and minerals, its imbalance can lead to diabetes [19]. Diabetes is one of the most prevalent chronic diseases worldwide. This research considers both clinical and non-clinical factors, such as insulin levels, glucose, BMI, smoking, stress, blood pressure, and dietary habits like junk food consumption. The study aims to highlight the impact of these factors on an individual's likelihood of developing diabetes. Various machine learning techniques, including Support Vector Machines (SVM), Decision Tree, Random Forest, and Logistic Regression, are explored for predicting and analyzing key features associated with

diabetes. The authors conduct a comparative analysis of these techniques to assess their performance when incorporating both clinical and non-clinical parameters.

Existing methods [20] for medical image feature extraction have proven insufficient in effectively addressing the challenges of early brain tumour detection. To overcome this limitation, a novel model leveraging the Inception-v3 convolution neural network has been proposed. This model enhances early identification of brain tumours by extracting and categorizing various features. Built upon the Inception-v3 architecture, it employs loss functions and the Adam Optimizer for hyper parameter optimization, and utilizes a softmax classifier for image classification into distinct categories. The model demonstrated a notable training accuracy of 99.02% and a validation accuracy of 89%, underscoring its potential in improving brain tumor detection.

The authors introduce an automated identification method that leverages deep learning and visual analysis technology [21]. Their approach classifies fundus images using a convolution neural network (CNN) based on the severity of diabetic retinopathy (DR). This strategy involves image modification, designing an efficient CNN architecture, and utilizing a large dataset for model training. The results demonstrate the effectiveness of this technology in accurately detecting residual drug traces.

The article presents a fuzzy distance-based ensemble approach integrating deep learning models for cervical cancer detection in Pap smear images [22]. The methodology employs three transfer learning models—Inception V3, MobileNet V2, and Inception ResNet V2—enhanced with additional layers to capture data-specific features. To combine the predictions from these models, the authors introduce a novel ensemble technique that minimizes the error between predicted and actual values. For cases with multiple predictions, three distance measures—Euclidean, Manhattan (City-Block), and Cosine—are computed for each class relative to the optimal solution. These measures are then defuzzified using the product rule to determine the final classification. Experimental results show that Inception V3, MobileNet V2, and Inception ResNet V2 individually achieved accuracies of 95.30%, 93.92%, and 96.44%, respectively. When the ensemble technique was applied, the overall performance improved to 96.96%, surpassing the individual models. The findings on three publicly available datasets confirm that the proposed model delivers competitive results compared to state-of-the-art approaches. This end-to-end classification method enhances cervical cancer detection in Pap smear images, aiding medical professionals in providing more effective treatment and improving the efficiency of the diagnostic process.

This study explores the effectiveness of machine learning and deep learning models in detecting heart murmurs from audio recordings. Utilizing the PhysioNet Challenge 2016 dataset [22], the authors compare traditional machine learning models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—with a Fully Convolution Neural Network (FCNN). While ensemble methods like Random Forest and AdaBoost improve robustness, they are still surpassed by deep learning techniques. The FCNN model, powered by artificial intelligence, outperforms all other models, achieving an accuracy of 0.99, a precision of 0.94, and a recall of 0.96. These findings underscore the potential of AI-driven cardiovascular diagnostics, as deep learning models demonstrate superior ability in recognizing complex patterns in heart sound data. The study suggests that deep learning models provide significant advantages in medical diagnostics, particularly in cardiovascular health, by offering scalable and highly accurate solutions for heart murmur detection.

This paper presents a systematic literature review of pneumonia detection techniques that incorporate transfer learning alongside other methodologies [23]. The review protocol is meticulously designed to identify recent research on pneumonia detection from the past five years. Following an extensive search process, 35 studies were selected for analysis. The review summarizes these papers, comparing the effectiveness of various pneumonia detection methods

based on their best-performing models. Additionally, the models are categorized into three primary approaches: deep learning methods, transfer learning techniques, and hybrid approaches. A comparative analysis of the top-performing models for pneumonia detection is also provided. The study concludes that while transfer learning demonstrates significant potential for enhancing pneumonia detection, further research is needed to refine these models for clinical use. Moreover, this review serves as a valuable resource for researchers by identifying existing research gaps in pneumonia detection techniques and suggesting directions for future advancements.

This study introduces a novel approach utilizing Deep Separable Convolution Neural Networks (DS-CNNs) to enhance Chronic Kidney Disease (CKD) prediction [24]. Using the Chronic Kidney Disease Dataset from Kaggle, the proposed model integrates DS-CNNs with advanced optimization techniques to improve predictive accuracy. DS-CNNs employ depth wise and point wise convolutions to enable efficient feature extraction and classification while maintaining computational efficiency. To further refine model performance, the Learning Rate Warm-Up with Cosine Annealing method is implemented, ensuring stable convergence and a controlled reduction in the learning rate. With an accuracy of 94.50%, the DS-CNN model surpasses traditional methods, offering superior predictive capabilities. These findings highlight the effectiveness of deep learning and optimization techniques in early CKD detection, presenting a promising tool for improved clinical decision-making.

The authors introduce and evaluate an innovative method for heart disease prediction by integrating deep learning models with bio inspired algorithms [25]. Deep learning techniques facilitate automatic feature extraction and the recognition of complex patterns from raw data, while bio inspired algorithms enhance optimization, improving model accuracy and generalization. Specifically, the cuckoo search algorithm and elephant herding optimization algorithm are utilized to fine-tune the architecture and hyper parameters of deep learning models, enabling an extensive exploration of different model configurations. This hybrid strategy efficiently combines the strengths of deep learning and bio inspired optimization, leading to the development of highly effective predictive models. Experimental results on benchmark heart disease datasets confirm the superiority of the proposed approach over traditional methods, demonstrating higher accuracy and robustness in predicting heart disease risk.

This study by the authors introduces an IoMT-enabled approach for lung disease detection and classification, leveraging deep learning techniques to analyze lung sounds [26]. The proposed method utilizes three datasets: the Respiratory Sound, the Coronahack Respiratory Sound, and the Coswara Sound. To benchmark performance, traditional machine learning models such as the Extra Tree Classifier and AdaBoost Classifier are employed. The Extra Tree Classifier achieved accuracies of 94.12%, 95.23%, and 94.21% across the datasets, while the AdaBoost Classifier demonstrated improvements with accuracies of 95.42%, 96.33%, and 94.76%. In contrast, the proposed deep neural network (DNN) outperformed these models, achieving accuracies of 98.92%, 99.33%, and 99.36% across the same datasets, highlighting its effectiveness in lung disease classification.

Imbalanced classification presents a significant challenge in early disease detection and diagnosis using machine learning, often leading to reduced accuracy due to the disproportionate distribution of positive cases compared to healthy individuals [27]. To enhance classification accuracy, this study proposes an architectural model that incorporates a modified Synthetic Minority Over-sampling Technique (SMOTE) with Minkowski distance and entropy-based weighting to determine the number of synthetic samples to generate. Additionally, feature selection is performed using a hybrid Particle Swarm Optimization-Grey Wolf Optimization (PSO-GWO) approach. The datasets used in the study are categorized based on feature count and total sample size into high, medium, and low-dimensional datasets. Six classification

algorithms are evaluated across datasets for diabetes, heart disease, and breast cancer. The final results demonstrate average accuracies of 74% for diabetes, 83% for heart disease, and 96% for breast cancer. The proposed approach effectively addresses class imbalances in medical datasets and outperforms traditional classification models, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest.

This study explores the use of a hybrid Autoencoder-LSTM (AE-LSTM) model to enhance the detection of Diabetic Nephropathy (DN) [28]. The Autoencoder (AE) component compresses clinical data while preserving essential features and reducing dimensionality, whereas the Long Short-Term Memory (LSTM) network captures temporal dependencies and sequential patterns, improving feature learning for early diagnosis. The dataset incorporates clinical and demographic variables, including age, sex, diabetes type, disease duration, smoking habits, and alcohol consumption. Implemented in Python, the proposed model demonstrates superior performance compared to conventional methods. The hybrid AE-LSTM model achieves an accuracy of 99.2%, marking a 6.68% improvement over Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression. These findings underscore the effectiveness of deep learning in accurately detecting DN at an early stage, providing a valuable tool for proactive disease management in diabetic patients.

This study presents an innovative approach to colon cancer diagnosis by integrating autoencoder - based feature selection, Capsule Networks (CapsNets), and histopathology images to address existing challenges [29]. CapsNets effectively capture spatial hierarchies in visual data, enhancing pattern recognition and classification accuracy. Autoencoders, when used for feature extraction, aid in dimensionality reduction, emphasize key features, and filter out noise, thereby improving overall model performance. The proposed method has achieved outstanding results, attaining an accuracy of 99.2%. The model's high accuracy in distinguishing malignant from non-malignant tissues underscores its ability to detect cancerous lesions with minimal errors. By combining autoencoders with Capsule Networks, this research marks a significant advancement in cancer detection, overcoming the limitations of conventional methods and providing a more reliable tool for early diagnosis.

This study introduces a versatile framework that utilizes a lightweight Convolution Neural Network (CNN) architecture to automate lung and colon cancer diagnosis in histopathological images across various diagnostic scenarios [30]. The commonly used LC25000 dataset, comprising 25,000 histopathological images classified into five categories—lung adenocarcinoma, lung squamous cell carcinoma, benign lung tissue, colon adenocarcinoma, and benign colonic tissue—serves as the foundation for this research. The study explores three diagnostic scenarios: (S1) differentiation between lung and colon samples, (S2) classification of benign and malignant images, and (S3) categorization into the five classes of the LC25000 dataset. The proposed model achieves exceptional performance across all scenarios, with accuracy, recall, precision, F1-score, and AUC exceeding 0.9947, 0.9947, and 0.9995, respectively. The lightweight CNN, with only 1.612 million parameters, proves to be highly efficient in automating lung and colon cancer diagnosis, outperforming several existing approaches.

This research presents an innovative chest X-ray classification framework that utilizes a fine-tuned VGG19 model (16 layers) enhanced with Contrast Limited Adaptive Histogram Equalization (CLAHE) for improved contrast, binary mask attention to emphasize abnormalities, and advanced data augmentation techniques for enhanced generalization [31]. A key feature of this approach is the implementation of a Probabilistic U-Net for lung segmentation, which isolates critical features while employing weighted masks to focus on pathological regions. To address class imbalance, computed class weights are incorporated to ensure fair learning. The proposed method achieves an impressive 95% accuracy, along with superior class-specific performance metrics, surpassing existing deep learning techniques. In

real-world healthcare applications, a test accuracy of 94.8% is obtained using various customized VGG19-based models, even without mask utilization, demonstrating the model's robustness and interpretability.

This study introduces the Hierarchical Reinforcement Learning with Multi-Expert Feedback (ME-RL) framework aimed at enhancing automated disease diagnosis within dialogue systems [32]. The framework employs a hierarchical structure with lower-level networks incorporating a reward model. In this setup, a discriminator, inspired by adversarial networks, evaluates the authenticity of symptom query sequences generated by the agent, providing rewards accordingly. Additionally, a large language model, reflecting human expert knowledge, assesses the appropriateness of the agent's current symptom queries, guiding the learning process of the policy network. This approach addresses data characteristic deficiencies and improves the policy's ability to utilize feature information, aligning the diagnostic process more closely with clinical practices. Experimental results demonstrate that the ME-RL framework achieves diagnostic success rates of 61.5% on synthetic datasets and 84.4% on real-world datasets, while also reducing the average number of dialogue turns required. These outcomes surpass those of conventional methods, indicating the framework's robust generalization capabilities.

This study [33] utilizes the Mask R-CNN semantic segmentation technique, incorporating a ResNet-50 backbone, to analyze CT scans of COVID-19 patients. The model is trained on an annotated dataset, improving its capability to accurately segment and delineate the lung parenchyma in CT images. Experimental results indicate that Mask R-CNN achieved a Dice Similarity Coefficient (DSC) of 93.4%, reflecting a high level of agreement between the segmented areas and clinically relevant regions. These findings underscore the effectiveness of the proposed approach in achieving precise lung tissue segmentation, facilitating quantitative assessments of lung impairments, and offering valuable insights for diagnosis and patient monitoring.

The study investigates the efficacy of combining various re sampling techniques with machine learning algorithms to improve prediction accuracy in imbalanced heart and lung disease datasets [34]. The authors integrate under sampling methods, such as Edited Nearest Neighbors (ENN) and Instance Hardness Threshold (IHT), with oversampling techniques like Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). These re sampling strategies are paired with classifiers including Decision Trees (DT), Random Forests (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM). Model performance is evaluated using metrics such as accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC). The results indicate that tailored re sampling significantly enhances machine learning model performance in healthcare settings. Notably, SVM combined with ENN under sampling markedly improves accuracy for lung cancer predictions, while SVM and RF with IHT achieve higher validation accuracies for both diseases. Random oversampling exhibits variable effectiveness across datasets, whereas SMOTE and ADASYN consistently enhance accuracy.

Glaucoma, a primary cause of irreversible blindness, is marked by characteristic changes in the optic disk. While manual assessments remain valuable, they are hindered by subjectivity, inconsistencies, and the significant time required for evaluation [35]. With the advent of artificial intelligence, machine learning (ML) and deep learning (DL) models have emerged as powerful tools for automated and improved optic disk analysis. Convolutional Neural Networks (CNNs) have been at the forefront of DL advancements, with novel architectures enhancing both specificity and sensitivity. Additionally, hybrid models that combine traditional ML techniques with DL approaches have demonstrated increased robustness and generalizability. ML and DL technologies hold transformative potential in glaucoma diagnosis and management, offering a balance of accuracy, efficiency, and reliability. As these models

continue to evolve by integrating larger datasets and multimodal imaging techniques, their clinical applications are expected to expand, fostering a stronger connection between technological progress and patient-centered care.

A neural network, a fundamental component of artificial intelligence, is designed to emulate the structure and function of the human brain [36]. It consists of interconnected nodes, or artificial neurons, organized into layers that process information sequentially. Each neuron receives inputs, performs computations, and transmits outputs to subsequent neurons. Through training on extensive datasets, neural networks adjust the weights of connections between neurons, enhancing their ability to recognize patterns and make decisions. This adaptability has led to significant advancements in fields such as speech recognition, computer vision, and natural language processing. This paper offers a concise overview of various machine learning paradigms, application areas, different types of neural networks, and their respective applications.

This research contributes to ongoing efforts in cardiovascular disease (CVD) detection by utilizing two powerful machine learning techniques: Multilayer Perceptron (MLP) and K-Nearest Neighbours (K-NN) [37]. The study employs publicly available data from the University of California Irvine repository, refining model performance by eliminating outliers and attributes with null values. Experimental results indicate that the MLP model achieves a superior accuracy of 82.47% and an area-under-the-curve (AUC) value of 86.41%, outperforming the KNN model. Therefore, the proposed MLP model is recommended as an effective solution for automated CVD detection.

The shift from traditional in-person classrooms to online learning environments has introduced challenges for both educators and students [38]. In virtual settings, instructors often concentrate on delivering content without real-time awareness of students' emotional states, potentially leading to decreased engagement and increased feelings of fatigue and disinterest among learners. To address this issue, a prototype system has been developed to assist educators in evaluating students' behaviours and emotions during online sessions. This system employs facial detection technology to identify features such as the eyes, nose, and mouth, enabling the differentiation and detection of multiple faces. By integrating a Convolution Neural Network (CNN) trained to predict emotional states through facial expressions, the system aims to provide insights into students' affective behaviours, thereby enhancing the effectiveness of online education.

The authors [39] introduce a novel diagnostic framework that integrates a one-dimensional convolution neural network (CONV1D) with an enhanced long short-term memory (LSTM) model. To assess data classification performance, confusion matrices are utilized within a deep learning algorithm. The proposed framework classifies data by incorporating key mechanical factors of bearings, such as sudden load variations, rotational speed, and operating temperature. This approach significantly enhances the model's training accuracy, achieving over 96.6%, with a percentage error of 23.29% after 50 iterations. Increasing the number of iterations could further improve accuracy, bringing it closer to 100% while reducing the error margin toward 0%.

Machine learning (ML) technology [40] identifies patterns linked to specific diseases by analyzing extensive datasets containing patient records, including diabetes, blood pressure, cholesterol levels, X-rays, MRIs, CT scans, imaging data, and genomic information. ML algorithms assess key symptoms to determine disease presence. The effectiveness of these models relies on the availability of sufficient data and the selection of relevant features for computation. The performance and fairness of an ML model depend on the chosen features used to diagnose a disease. Selecting too many or too few features can lead to under fitting or over fitting, impacting prediction accuracy. Ensuring all essential attributes are included while avoiding irrelevant ones is a challenging yet crucial aspect of model development. Omitting a

key attribute or incorporating unrelated features can distort the model's outcomes. This research investigates the impact of feature selection and bias in ML disease prediction by employing Support Vector Machine (SVM) and Logistic Regression (LR) algorithms. The study highlights the importance of careful feature selection to improve model reliability and fairness in disease diagnosis.

Multi-objective optimization algorithms aim to identify optimal trade-offs among conflicting objectives [41]. Among these, Particle Swarm Optimization (PSO)-based multi-objective algorithms are notably prominent. This study conducts a comparative analysis of Machine Learning (ML) algorithms within the context of Fuzzy Multi-objective Evolutionary Algorithms (FMOEAs), utilizing classification algorithms from the WEKA software suite and datasets from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository. The FMOEA approach is assessed using a benchmark medical dataset to evaluate its effectiveness. The findings indicate that the FMOEA method effectively balances the trade-off between accuracy and interpretability, offering valuable insights for researchers and practitioners in selecting appropriate algorithms for classification tasks.

This study compares four machine learning classifiers—Decision Tree, K-Nearest Neighbour, Linear Discriminant Analysis, and Random Forest—using evaluation metrics such as accuracy, recall, specificity, and F1 score [42]. The dataset includes patients diagnosed with chronic obstructive pulmonary disease, asthma, bronchiectasis, as well as healthy individuals. Results indicate that the Random Forest classifier outperforms the others, achieving an impressive accuracy of 99.72%, along with perfect recall, specificity, and F1 scores. The findings highlight the potential of computer-based systems as reliable decision-making tools for pulmonary disease classification, particularly in resource-constrained environments.

This study introduces a multi-criteria decision-making (MCDM) framework to evaluate machine learning techniques (MLTs) for diabetes prediction [43]. Initially, three MCDM methods—Weighted Sum Model (WSM), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) are been applied. Subsequently, a fusion approach consolidates these individual rankings to determine a final, comprehensive ranking of the MLTs. The framework's effectiveness is validated by assessing the performance of ten MLTs on the Pima Indian Diabetes dataset using eight evaluation metrics. The consolidated MCDM rankings suggest that logistic regression is the most suitable technique for diabetes prediction modeling.

This study [44] introduces a novel deep learning framework, Efficient DenseNet, for classifying diabetic retinopathy severity levels in retinal images. The proposed model is a thin-layered Efficient DenseNet architecture with fewer trainable parameters, achieving superior classification accuracy compared to other deep learning models. The framework incorporates an automatic pre-processing module, followed by data augmentation and high-level feature extraction using the Efficient DenseNet model and classifier. Trained and tested on 13,000 retinal fundus images from a diabetic retinopathy database, the model, when combined with a k-nearest neighbour classifier, achieves an impressive classification accuracy of 98.40%.

This study [45] focuses on developing an artificial intelligence module to recognize abnormal tension patterns in textile weaving, addressing the limitations of traditional manual methods. Utilizing Long Short-Term Memory (LSTM) recurrent neural networks, the model is trained and validated on five common abnormal tension patterns. The approach integrates plug-in modules and edge computing within deep learning frameworks, allowing implementation without altering the existing system architecture. Extensive experiments were conducted to optimize model parameters. The results demonstrate an average recognition rate of 97.12% for abnormal tension patterns, with an average computation time of 46.2 milliseconds per sample, indicating that the system meets practical performance requirements.

This study [46] investigates the feasibility of utilizing machine learning techniques to determine dietary intake types based on laboratory test data. Specifically, seven distinct

machine learning algorithms were applied to analyze complete blood count (CBC) data, encompassing 15 variables from 44 laboratory animals subjected to either a standard diet or a high-fat diet. The evaluation metrics considered include accuracy, precision, recall, and F1-score. The results demonstrate that these techniques effectively identify the type of diet, achieving accuracy rates exceeding 88%. This approach presents a promising alternative to assist health professionals in decision-making processes.

An Ant Colony System (ACS)-based segmentation method [47] is introduced for segmenting the optic disc and macula in retinal images. This approach utilizes intensity (gray-level) and colour as distinguishing features to accurately localize and segment these structures. The proposed ACS-based segmentation technique achieves accuracy levels comparable to those attained by supervised pixel classification algorithms, demonstrating its effectiveness in retinal image analysis.

Arrhythmias manifest in various forms, with some presenting no symptoms at all [48]. When symptoms do occur, individuals may experience palpitations or a sensation of skipped heartbeats. In more severe cases, symptoms such as dizziness, fainting, shortness of breath, or chest pain may arise. While many arrhythmias are benign, certain types can lead to severe complications, including stroke, heart failure, or even cardiac arrest. Arrhythmias impact millions of people worldwide, contributing significantly to cardiovascular-related fatalities. Sudden cardiac death accounts for nearly half of all cardiovascular deaths, making up about 15% of global mortality, with ventricular arrhythmias responsible for approximately 80% of these cases. Although arrhythmias can occur at any age, they become increasingly common with aging.

Diabetic retinopathy (DR) is a leading cause of vision impairment, necessitating regular eye screenings for timely treatment and prevention of vision loss [49]. The global rise in DR cases has increased the workload on specialists and diagnostic equipment. Fundus imaging plays a crucial role in effective retinal assessments. To enhance DR detection from retinal images, a deep learning approach has been proposed, integrating four key techniques: image augmentation, contrast-limited adaptive histogram equalization (CLAHE), convolution neural networks (CNNs) with transfer learning, and ensemble classification. This method achieved notable performance metrics, including an accuracy of 93%, precision of 95%, and recall of 96%, demonstrating greater stability compared to other approaches.

Medical image noise can significantly affect both diagnostic accuracy and image clarity, making it challenging to define object boundaries essential for precise assessments [50]. Implementing denoising techniques can greatly enhance image quality, reduce diagnostic errors, and improve treatment accuracy. This study introduces a method for mitigating various types of noise, including Gaussian noise, salt-and-pepper noise, speckle noise, and ring artifacts, in medical imaging modalities such as MRI, CT, and chest X-rays. The approach integrates adaptive convolution neural networks (CNNs) with guided image filtering to improve image quality and applies deep learning-based figure-ground segmentation. The proposed method is rigorously tested on chest X-ray and MRI/CT images across different noise conditions, with performance comparisons against existing techniques using well-established statistical metrics. The results confirm that this approach achieves superior performance, consistently yielding the highest metric values.

The referenced article introduces an innovative ensemble approach utilizing deep learning models for the detection of cervical cancer in Pap smear images [51]. The methodology incorporates three transfer learning models—Inception V3, MobileNet V2, and Inception ResNet V2—each augmented with additional layers to capture data-specific features. To combine the outputs of these models, a novel ensemble technique is employed, focusing on minimizing the error between observed outcomes and ground-truth labels. For instances with multiple predictions, three distance metrics—Euclidean, Manhattan (City-Block), and

Cosine—are calculated for each class relative to their optimal solutions. These distance measures are then defuzzified using the product rule to determine the final predictions. Individually, the models achieved accuracies of 95.30% for Inception V3, 93.92% for MobileNet V2, and 96.44% for Inception ResNet V2. Upon applying the proposed ensemble method, the overall performance improved to 96.96%, surpassing the results of the individual models. The experimental findings, based on three publicly available datasets, demonstrate that the proposed model delivers competitive results compared to existing state-of-the-art methods. This approach offers an end-to-end classification technique for detecting cervical cancer from Pap smear images, potentially aiding medical professionals in providing better treatment and enhancing the efficiency of the diagnostic process.

Ruhul Amin. et. Al. [52] employ integrated projection-based statistical feature extraction techniques to minimize data redundancy in the classification of chronic liver disease. Utilizing the Indian Liver Patient Dataset (ILPD) from the University of California, Irvine (UCI) repository—which includes 583 patient records (416 diagnosed with liver disease and 167 without)—the study incorporate multiple machine learning models: logistic regression (LR), random forest (RF), K-nearest neighbour (KNN), support vector machine (SVM), multilayer perceptron (MLP), and an ensemble voting classifier. The proposed methodology achieved an accuracy of 88.10%, surpassing previous studies by 0.10–18.5%.

Summary of Literature Review are as Follows:-

Table 1

Methods used and objective of paper by authors in Literature Review

| Paper No | Methods | Key Finding |
|----------|--|--|
| 1 | Machine Learning models | Health care(oral) disease prediction with an objective to recovery |
| 2, 4, 6 | Machine Learning models | Health care disease prediction with an objective to recover. |
| 5 | Machine Learning models using gene expression data | Lung cancer detection & recovery. |
| 7 | Ensemble Techniques | Heart disease detection & recovery |
| 9 | Segmentation & deep learning techniques | Health care disease prediction with an objective to recovery |
| 10 | CNN and LSTM models with ECG signals | Health care disease prediction with an objective to recovery |
| 11 | CNN models | Heart disease prediction with an objective to recovery |
| 12 | Machine Learning models | Breast Cancer detection & recovery. |
| 13 | LSTM, CNN models | Breast Cancer detection & recovery |
| 14 | Segmentation | Lung cancer prediction using CT scan images |
| 15 | CNN + RELU algorithms | BraTs2020 data set used for detection brain tumor. |
| 16 | Deep learning models | Cardiac abnormalities detection. |
| 17 | Machine learning models | Cancer prediction |
| 18 | MobileNet, CNN models viz. VGG16, Xception, Resnet50 | Brain tumor prediction. |
| 19 | Machine Learning models | Diabetics prediction |
| 20 | InceptionV3 model | Brain tumor prediction |

| | | |
|--------|--|--|
| 21 | CNN | Diabetics Retinopathy prediction |
| 22, 25 | Deep learning models | Heart disease prediction |
| 23 | Deep learning, transfer learning, hybrid models | Pneumonia detection |
| 24 | Deep CNN | Chronic kidney disease prediction |
| 26 | Deep learning models | Lung disease detection |
| 28 | Hybrid Autoencoder & LSTM | Diabetics prediction |
| 29 | Autoencoder with capsule Network | Colon cancer prediction |
| 30 | Lightweight CNN | Colon cancer prediction |
| 31 | CNN | Chest disease prediction |
| 33 | Segmentation mask RCNN with ResNet 50 | Lung disease detection |
| 35 | Machine learning & Deep learning models | Lung disease detection& recovery |
| 37 | Machine learning models | Cardiovascular disease prediction |
| 38 | CNN and LSTM | Prediction of students' emotion in class |
| 42 | Machine learning models | Pulmonary disease prediction |
| 43 | Reinforcement based learning | Diabetics prediction |
| 44 | Deep learning models using DenseNet | Retinopathy prediction |
| 45 | Deep learning with LSTM | Abnormal tension prediction |
| 46 | Machine learning models | Diabetics prediction |
| 48 | Machine learning models | Arrhythmias detection |
| 49 | CNN | Diabetics Retinopathy prediction |
| 50 | Deep learning models with segmentation using chest x ray images | for detection of diabetic retinopathy |
| 51 | Deep learning models Inception v3, MobileNet v2, Inception ResNet v2 | Cancer detection |
| 52 | Machine learning models | Liver Patient Dataset |

3. Motivation.

A lot of research work ([1]-[52]) have been done in the area of healthcare prediction with an objective to detect sickness of various organs. Machine learning algorithms have also been proposed([1], [2], [4]-[6], [12], [17], [19], [37], [42], [46], [48], [52]). Deep Learning models have been proposed in {[9], [16], [22], ([24]-[26]), [44], [45], [50], [51]}, Segmentation has been done in ([9], [14], [33]). However, no author has worked on the same data set and not evaluated several evaluation measures. That is the reason for this proposed work which has been written in this paper. Here an effort is being made to work using machine learning models and cross validation estimation learning techniques on lever disease data set with an objective to select a particular model suitable for the data set. After selecting suitable model, it is necessary to find out the disease affected persons so that in course of time, the concerned person may be recovered with application of proper medicine and food and other items as the case may be.

4. Data set.

The lever data set has been collected from UCI Machine Repository ([56]). The

data set comprises of 7 attributes. These are furnished below:-

Table 2
Lever Data Set

| No | Attribute Code | Attribute Information |
|----|----------------|--|
| 1 | mcv | mean corpuscular volume |
| 2 | alkphos | alkaline phosphotase |
| 3 | sgpt | alamine aminotransferase |
| 4 | sgot | aspartate aminotransferase |
| 5 | gammagt | gamma-glutamyltranspeptidase |
| 6 | drinks | number of half-pint equivalents of alcoholic beverages drunk per day |
| 7 | selector | field used to split data into two sets viz. healthy or diseased. |

5. Methodology.

5.1. Random Forest.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Random forests are an ensemble learning method for classification, regression. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Algorithm

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 and Step 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

5.2. Decision tree

Decision Trees are flowchart-like tree structures of all the possible solutions to a decision, based on certain conditions. It is called a decision tree as it starts from a root and then branches off to a number of decisions just like a tree. The tree starts from the root node where the most important attribute is placed. The branches represent a part of entire decision and each leaf node holds the outcome of the decision.

The best attribute or feature is selected using the Attribute Selection Measure(ASM). The attribute selected is the root node feature. Attribute selection measure is a technique used for the selecting best attribute for discrimination among tuples. It gives rank to each attribute and the best attribute is selected as splitting criterion. The most popular methods of selection are: (1) Entropy(2). Information Gain (3). Gain Ratio (4). Gini Index

5.2.1. Entropy. Entropy is the randomness in the information being processed. It measures the purity of the split.

This algorithm compute the entropy with the following formula: $-(p \log_2(p)) - (q \log_2(q))$ p is the probability of success or the number of positive cases. q is the probability of failure or the number of negative cases.

5.2.2. Information gain.

Information gain is used to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned. Information gain tells us how important a given attribute of the feature vectors is. The ordering of attributes in the nodes of a decision tree can be decided with the help of this feature.

5.2.3. Gain Ratio.

The gain ratio means the share of profit gained by a partner with some reconstitution of the firm. This gaining ratio is caused by the reconstitution which generally happens due to the exit or death of any existing partner.

5.2.4. Gini Index

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower Gini index should be preferred.

The Formula for the calculation of the Gini Index is given below:-

Giniindex = $1 - \sum p_j^2$ where p_j is the probability of an object being classified to a particular class.

5.3. KNN(K Nearest Neighbors) algorithm: It is one of the simplest and widely used classification algorithms in which a new data point is classified based on similarity in the specific group of neighbouring data points.

Algorithm.

Step 1: Select the value of K neighbours

Step 2: Find the K nearest data point for our new data point based on euclidean distance

Step 3: Among these K data points count the data points in each category

Step 4: Assign the new data point to the category that has the most neighbours of the new data point.

5.4. Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm. It can be used in both classification and regression problems. Inherently, it is a discriminative classifier. Given a set of labelled data points, an SVM tries to separate the data points into different output classes. It does so by finding an optimal hyper plane that distinctly classifies the data points into an N-dimensional space(N - the number of features).

Support vectors are those two data points supporting the decision boundary (the data points which have the maximum margin from the hyper plane). Support Vector Machine(SVM) always makes an effort to those two data points from different classes that are the closest to each other. These support vectors are the keys to draw an optimal hyper plane by SVM. In SVM, the set of input and output data are treated as vectors. This is because when the data is a higher dimensional space (more than two dimensions), the classes cannot be represented as single data points, so they must be represented as vectors.

Non-linearly separable data can be separated by taking them into a higher dimension. It is necessary to map the data into just one dimension higher. Kernel Tricks are functions that apply on some complex mathematical operations on the lower-dimensional data points and convert them into higher dimensional space. Then finds out the process of separating the data points based on the labels and outputs. Common kernels are Linear Kernel, Polynomial Kernel, Radial Basis Function(RBF) of RBF Kernel, Sigmoid Kernel, Gaussian Kernel.

5.5. Linear Regression.

Linear regression tries to model the relationship between two variables by fitting a linear equation to the available data. One variable is taken as an explanatory variable, and the other is called as dependent variable. The form of the regression equation with one dependent and one independent variable is defined as $y = c + b*x$, where y is considered as dependent variable score, c = constant, b = regression coefficient, and x = an independent variable

Which takes a number of values.

5.6. Confusion Matrix

The confusion matrix is a matrix which is used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. Some features of Confusion matrix are given below:

For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.

The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.

Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

The following table shows Confidence Matrix:

| Total Predictions | Actual: No | Actual: Yes |
|--------------------------|---------------------|---------------------|
| Predicted: No | True Negative (TN) | False Positive (FP) |
| Predicted: Yes | False Negative (FN) | True Positive (TP) |

The above table has the following cases:

True Negative (TN): Model has given prediction No, and the real or actual value was also No.

True Positive (TP): The model has predicted yes, and the actual value was also true. It is called Type-I error.

False Negative (FN): The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

False Positive (FP): The model has predicted Yes, but the actual value was No.

It evaluates the performance of the classification models, predictions work on test data, and tells how good our classification model is.

It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.

With the help of the confusion matrix, calculate the different parameters for the model, such as accuracy, precision, etc. can be calculated.

5.7. Classification Report.

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of the classification model. It provides a better understanding of the overall performance of our trained model.

Precision: Precision is defined as the ratio of true positives to the sum of true and false positives.

Recall: Recall is defined as the ratio of true positives to the sum of true positives and false negatives.

F1 Score: The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

Support: Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

5.8 Cross - Validation Methods.

5.8.1. K fold Cross-Validation

To implement k-fold cross-validation, the original dataset is divided into k number of partitions. The holdout method is then performed k number of occasions, each time using a different partition as the testing set, while the remaining partitions are used for training. This repeated process helps to obtain a more reliable and robust evaluation of the model's performance by leveraging a larger amount of data for testing and training purposes.

5.8.2. Cross Validation Techniques using estimator values.

A random forest is an ensemble learning method that combines the predictions from multiple decision trees to produce a more accurate and stable prediction. It is a type of supervised learning algorithm that can be used for both classification and regression tasks.

For a particular value of estimators, the number of decision trees will be formed in a particular machine learning models.

5.8.3. Stratified K-fold cross Validation.

Stratified K Fold used when just random shuffling and splitting the data is not sufficient, and it is necessary to have correct distribution of data in each fold. In case of regression problem folds are selected so that the mean response value is approximately equal in all the folds. In case of classification problem folds are selected to have same proportion of class labels. Stratified K Fold is more useful in case of classification problems, where it is very important to have same percentage of labels in every fold.

5.8.4. Leave-One-Out Cross-Validation (LOOCV):

In this method, the model is trained on the entire dataset but leaves out only one data point of the available dataset and then iterates for each data point. It has some advantages as well as disadvantages. An advantage of using this method is that all data points can be used and hence it is low bias. The major drawback of this method is that it leads to higher variation in the testing model as one data point is used for testing. If the data point is an outlier it can lead to higher variation.

5.8.5. Hold Out Method.

The holdout method is a basic Cross Validation approach in which the original dataset is divided into two discrete segments: Training Data -this set is used to fit and train the model. Test Data - This set is used to evaluate the model. In the majority of cases, the size of the training dataset is typically much larger than the test dataset. Therefore, a standard holdout method split ratio is 70:30 or 80:20. Furthermore, the overall dataset is randomly rearranged before dividing it into the training and test set portions using the predetermined ratio.

5.8.5.1. Lasso Regression.

Lasso regression, short for Least Absolute Shrinkage and Selection Operator, is a form of linear regression that incorporates regularization. It does this by adding a penalty term to the loss function, which is based on the absolute values of the coefficients. This technique is especially valuable when dealing with datasets that have a large number of features, as it can effectively perform feature selection by reducing the coefficients of less relevant variables to zero.

5.8.5.2. Ridge regression

Ridge regression, also called Tikhonov regularization, is a method used to handle multiple regression problems affected by multi collinearity. In cases of multi collinearity, ordinary least squares estimates remain unbiased but can have high variance, making them unreliable. Ridge regression addresses this by introducing a small amount of bias to the estimates, which helps lower their standard errors. This approach is especially helpful when the predictor variables are strongly correlated with each other.

5.8.6. Recursive Feature Elimination.

Recursive Feature Elimination is a feature selection method to identify a dataset's key features. The process involves developing a model with the remaining features after repeatedly removing the least significant parts until the desired number of features is obtained.

6. Contribution.

Proposed Flow of Work

Step 1. Data Collection. Lever data set Data have been collected from [56]

Step 2. The dataset containing relevant information for the prediction of disease of lever have been taken.

Step 3. The data have been entered accurately and completely.

- Step 4. Data Pre-processing: Cleaning of data and Removal of Outlier have been done.
- Step 5. Taking care of missing data by input certain concerned data or removal of that data based on the nature and quantity of missing values.
- Step 6. Cleaning the data by tackling inconsistencies, errors, and anomalies.
- Step 7. Detection and removal of outliers that may affect the analysis.
- Step 8. Implementation of Machine Learning models and Performance Evaluation:
- Step 9. Under machine learning classifier models, Random forest classifier, Decision Tree classifier, Support Vector Machine (Linear), Logistic Regression, Gaussian Naïve Bayes, K Nearest Neighbours have been used.
- Step 10. Train the classifiers using the pre-processed data.
- Step 11. Evaluate the performance of each classifier using evaluation metrics like accuracy, precision, recall, and F1 score.
- Step 12. Application of cross validation Methods and Performance Evaluation:
- Step 13. Cross validation methods include K fold cross validation, cross validation using estimated values, stratified cross validation, leave one out cross validation, hold out method, recursive feature elimination.
- Step 14. Evaluate K fold cross validation methods based on root mean square error value using a number of machine learning models.
- Step 15. Evaluate cross validation methods based on root mean square error value using a number of machine learning models for a number of estimated values.
- Step 16. Evaluate using a number of machine learning models under stratified K-fold cross Validation method.
- Step 17. Evaluate a number of machine learning models using Leave-One-Out Cross-Validation method using the value of accuracy.
- Step 18. Evaluate a number of machine learning models using hold out method in terms of the value of accuracy
- Step 19. Evaluate a number of machine learning models under the method of Recursive Feature Elimination using the value of accuracy.
- Step 20. Evaluate the performance of each cross validation techniques and compare the results.
- Step 21. Compare and analyse the performance of all the different techniques employed in previous steps (Machine Learning models and cross validation techniques).
- Step 22. Identify the most accurate and reliable method for predicting the lever disease based on the results as obtained.

6.1. Classification.

6.2. Application of Machine Learning models.

The application of machine learning models has been applied on input data [56]. The application of machine learning models has been furnished in figure 1.

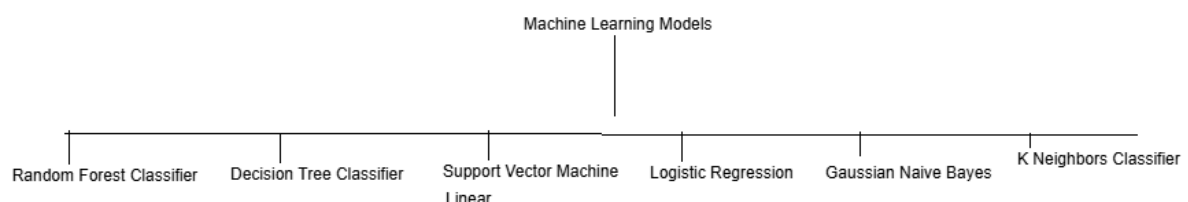


Figure 1

6.2.1. Random Forest.

Input data set as available has to be applied to random forest algorithm. It has used 10 estimators that means 10 decision trees have been constructed and finally the average of these tree values has to be taken. The value of accuracy has been found as 88.89 % for criterion as gini index as well as for entropy.

6.2.2. Decision Tree.

Input data set as available has to be applied to decision tree algorithm. Criterion as gini index has been used. The value of accuracy has been found as 89 %.

6.2.3.KNN(K Nearest Neighbours) algorithm:

Input data set as available has to be applied to KNN(K Nearest Neighbours algorithm. Number of neighbours has been used as 5. Distance function is used as 'minkowski'. The value of accuracy has been found as 77.77%.

6.2.4. Support Vector Machine algorithm.

Input data set as available has to be applied to support vector machine algorithm. The value of accuracy has been found as 91.66 % based on kernel function as linear. The value of accuracy has been found as 94.44 % based on kernel function as radial basis.

6.2.5. Linear Regression.

Input data set as available has to be applied to logistic regression algorithm. The value of accuracy has been found as 100 %.

6.2.6. Logistic Regression.

Input data set [11] has to be applied to logistic regression algorithm. The value of accuracy has been found as 86.66 %.

6.2.7. Gaussian Naïve Bayes Algorithm.

Input data set [11] has to be applied to Gaussian Naïve Bayes algorithm. The value of accuracy has been found as 94.44 %.

6.2.8. Gradient Boosting Classifier Algorithm.

Input data set [11] has to be applied to Gradient Boosting Classifier algorithm. The value of accuracy has been found as 94.44 %.

6.3. Application of Cross Validation Techniques.

The application of cross validation Techniques has been furnished in figure 2.

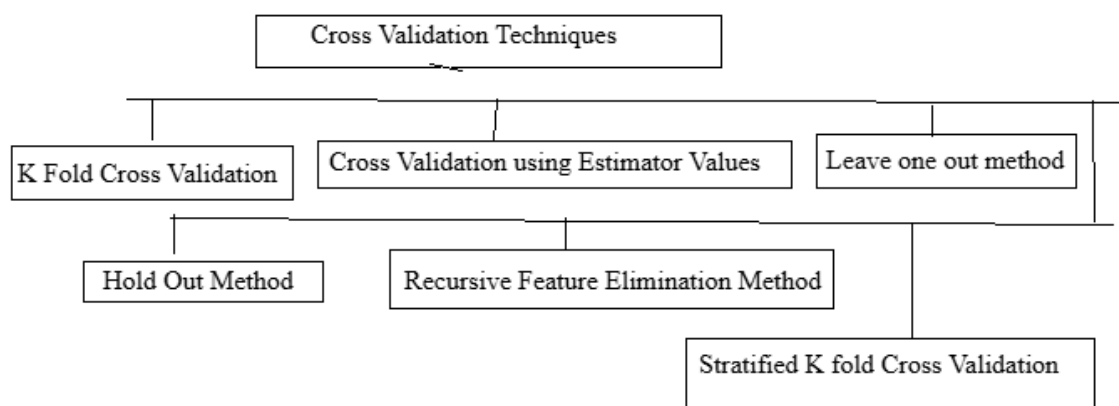


Figure 2

6.3.1. K fold Cross-Validation.

The value of K fold is chosen as 10. The methods of linear regression, decision tree regressor, random forest regressor, decision tree regressor have been applied based on this value of K fold.

6.3.2. Cross Validation Techniques using estimator values.

For the value of estimators is 10, the number of decision trees in the Random Forest is assigned as 10. Here the estimator values are chosen as 50, 100, 150, 200, 250, 300, 350 respectively. Therefore, the number of decision trees is same as estimator value.

6.4. Leave one out method.

In this method training is performed on the whole dataset but leaves only one data-point of the available dataset and then iterates for each data-point. Here the model is trained on $n-1$ samples and tested on the one omitted sample. This process is repeated for each data point in the dataset.

6.5. Hold out method.

Initially decision Tree classifier has been applied using different training and test size and the corresponding accuracy has been computed. It has been found that the accuracy is maximum for training / testing data size as 0.7 / 0.3

6.6. Recursive Feature Elimination.

Recursive Feature Elimination is a feature selection method to identify a dataset's key features. The process involves developing a model with the remaining features after

6.7. Training and Test data.

The number of training data has been used as 70% and that of test data as 30% for the above models for better performance of the models.

7. Results.

The comparative study of all machine learning models on the basis of accuracy have been furnished in table 3 as furnished below. The classification report, confusion matrix of all the models have been furnished in table

4, table 5 respectively.

Table 3
Machine Learning models versus accuracy

| No | Name of Machine Learning Model | Accuracy(%) |
|----|---|-------------|
| 1. | Random Forest Algorithm | 88.89 % |
| 2. | Decision Tree Algorithm | 89 % |
| 3. | KNN(K-Nearest Neighbor) Algorithm | 77.77 % |
| 4. | Support Vector Machine with Linear Kernel | 100 % |
| 5. | Linear Regression Algorithm | 100 % |
| 6. | Gaussian Naïve Bayes Algorithm | 94.44 % |
| 7. | Logistic Regression | 86.66 % |
| 8. | Support Vector Machine with Radial Basis Kernel | 94.44 % |
| 9. | Gradient Boosting Classifier | 94.44 % |

Table 4
Classification Report Item wise based on Machine Learning Models

| Model | Precision | Recall | f1-score | Support |
|-----------------------------|-----------|--------|----------|---------|
| Random Forest 0 | 0.91 | 0.91 | 0.91 | 11 |
| Random Forest 1 | 0.86 | 0.86 | 0.86 | 7 |
| Decision Tree 0 | 1.00 | 0.82 | 0.9 | 11 |
| Decision Tree 1 | 0.78 | 1.00 | 0.88 | 7 |
| KNN Classifier 0 | 0.82 | 0.82 | 0.82 | 11 |
| KNN Classifier 1 | 0.71 | 0.71 | 0.71 | 7 |
| SVM(Kernel=Linear) 0 | 1.00 | 1.00 | 1.00 | 11 |

| | | | | |
|---|-------|------|------|----|
| SVM(Kernel=Linear) 1 | 1.00 | 1.00 | 1.00 | 7 |
| Linear Regression 0 | 1.00 | 1.00 | 1.00 | 11 |
| Linear Regression 1 | 1.00 | 1.00 | 1.00 | 7 |
| Gaussian Naïve Bayes Algorithm 0 | 0.92 | 1.00 | 0.96 | 11 |
| Gaussian Naïve Bayes Algorithm 1 | 1.00 | 0.86 | 0.92 | 7 |
| Logistic Regression 0 | 0.90 | 0.90 | 0.90 | 11 |
| Logistic Regression 1 | 0.80 | 0.80 | 0.80 | 7 |
| SVM(Kernel=Radial Basis) 0 | 0.92 | 1.00 | 0.96 | 11 |
| SVM(Kernel=Radial Basis) 1 | 1.00 | 0.86 | 0.92 | 7 |
| Gradient Boosting Classifier 0 | 0.92 | 1.00 | 0.96 | 11 |
| Gradient Boosting Classifier | 11.00 | 0.86 | 0.92 | 7 |

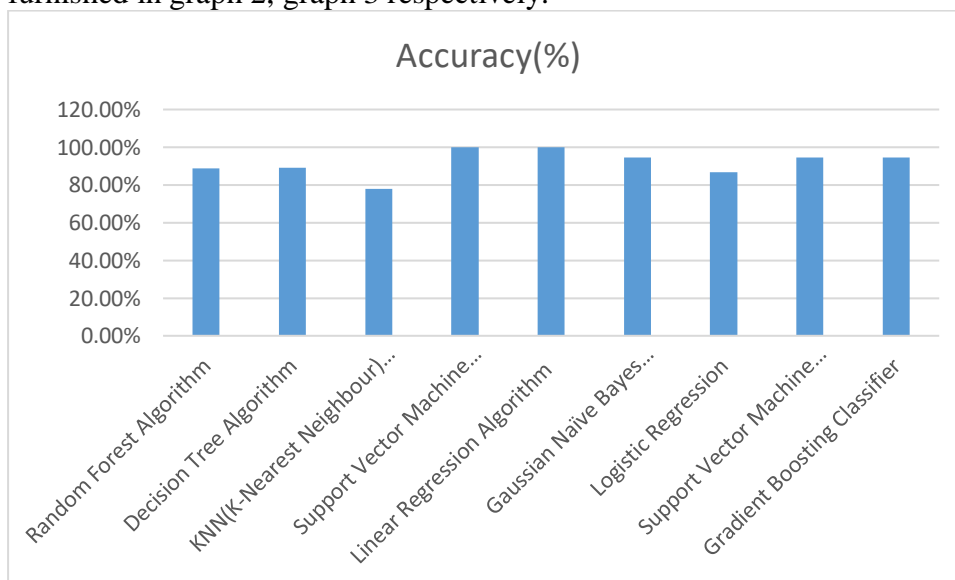
Table 5
Confusion Matrix based on Machine Learning Models

| Model Item | True Positive | False Positive | False Negative | True Negative |
|-------------------------------|----------------------|-----------------------|-----------------------|----------------------|
| Random Forest 0 | 10 | 1 | 1 | 6 |
| Random Forest 1 | 6 | 1 | 1 | 10 |
| Decision Tree 0 | 9 | 2 | 0 | 7 |
| Decision Tree 1 | 7 | 0 | 2 | 9 |
| KNN Classifier 0 | 9 | 2 | 2 | 5 |
| KNN Classifier 1 | 5 | 2 | 2 | 9 |
| SVM(Kernel=Linear) 0 | 11 | 0 | 0 | 7 |
| SVM(Kernel=Linear) 1 | 7 | 0 | 0 | 11 |
| Linear Regression 0 | 10 | 1 | 1 | 6 |
| Linear Regression 1 | 6 | 1 | 1 | 10 |
| Gaussian Naïve Bayes 0 | 11 | 0 | 1 | 6 |

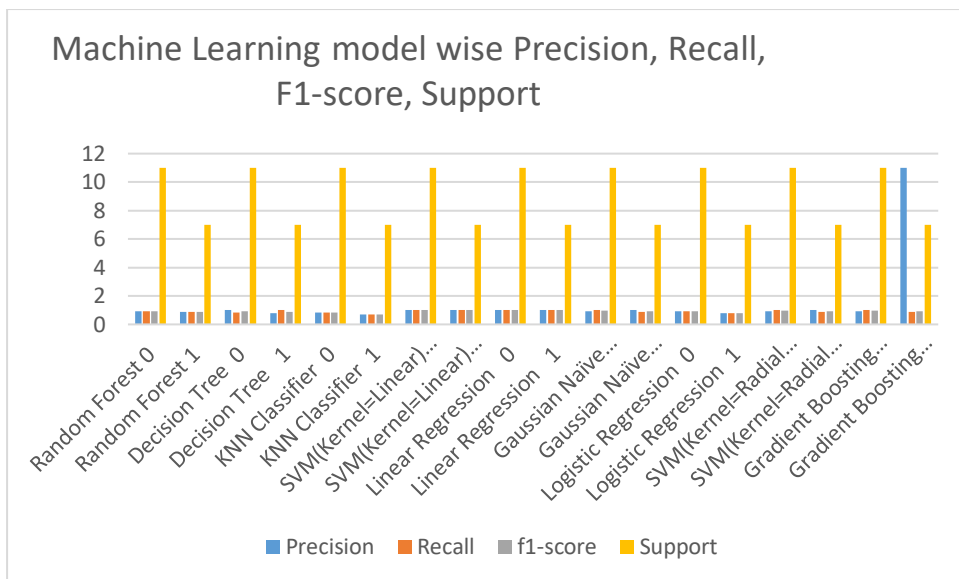
| | | | | |
|-------------------------------|----|---|---|----|
| Gaussian Naïve Bayes 1 | 6 | 1 | 0 | 11 |
| Logistic Regression 0 | 10 | 1 | 1 | 6 |
| Logistic Regression 1 | 6 | 1 | 1 | 10 |
| Gradient Boosting 0 | 11 | 0 | 1 | 6 |
| Gradient Boosting 1 | 6 | 1 | 0 | 11 |

The preference of model has to be decided on the more value of accuracy. From table 2 it has been observed that the accuracy of support vector machine with linear kernel, linear regression is 100.00 % which is the maximum value among all models. Therefore, support vector machine with linear kernel, linear regression has to be preferred. Considering theoretical concept, support vector machine with linear kernel has to be considered as compared to linear regression.

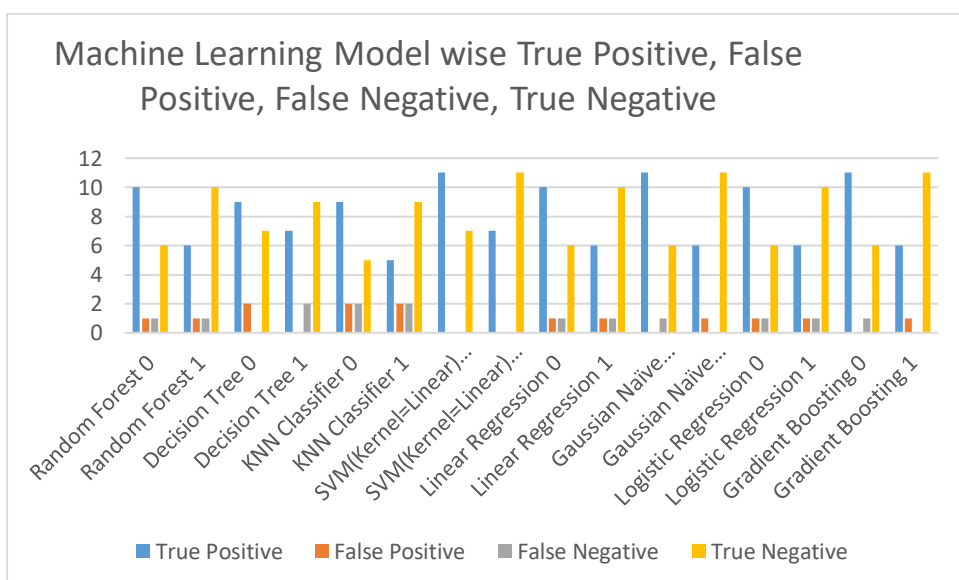
The Change of values of accuracy has been furnished in graph named graph 1. The change of values of precision, recall, f1-score (classification report) and confusion matrix has been furnished in graph 2, graph 3 respectively.



Graph 1



Graph 2



Graph 3

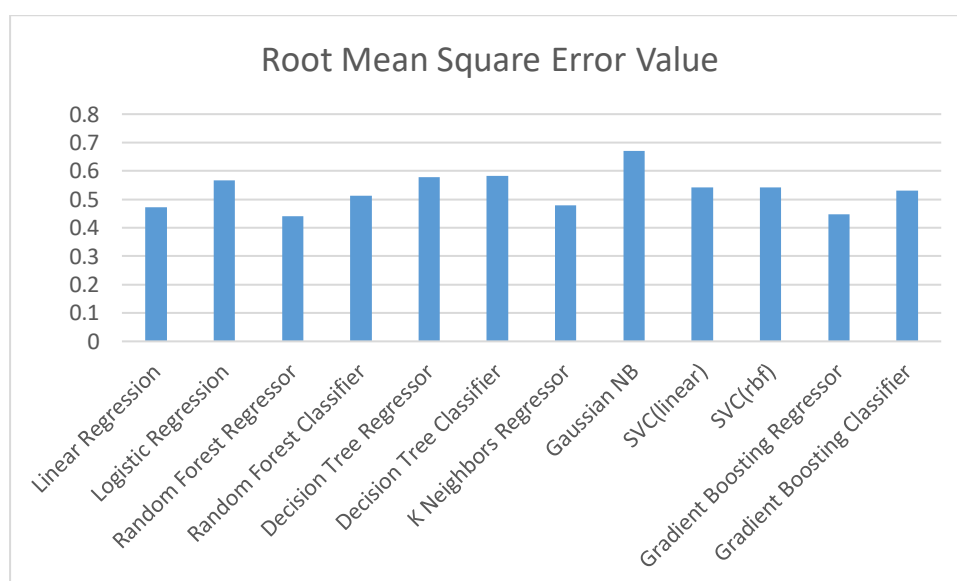
K-fold Cross Validation.

The root mean square error values based on machine learning models using K-fold Cross Validation has been furnished in table 6.

Table 6**Root Mean square error value versus Machine Learning Models**

| Machine Learning Models | Root Mean Square Error Value |
|------------------------------|------------------------------|
| Linear Regression | 0.4721 |
| Logistic Regression | 0.5676 |
| Random Forest Regressor | 0.4417 |
| Random Forest Classifier | 0.5136 |
| Decision Tree Regressor | 0.5776 |
| Decision Tree Classifier | 0.5826 |
| K Neighbors Regressor | 0.4789 |
| Gaussian NB | 0.6717 |
| SVC(linear) | 0.5413 |
| SVC(rbf) | 0.5413 |
| Gradient Boosting Regressor | 0.4479 |
| Gradient Boosting Classifier | 0.5301 |

From table 1 it is observed that Root Mean Square Error Value for Random Forest Regressor is 0.44, which is the minimum among other models. Therefore, Random Forest Regressor has to be selected. The Change of values of root mean square value based on machine learning models has been furnished in graph named graph 4.



Graph 4

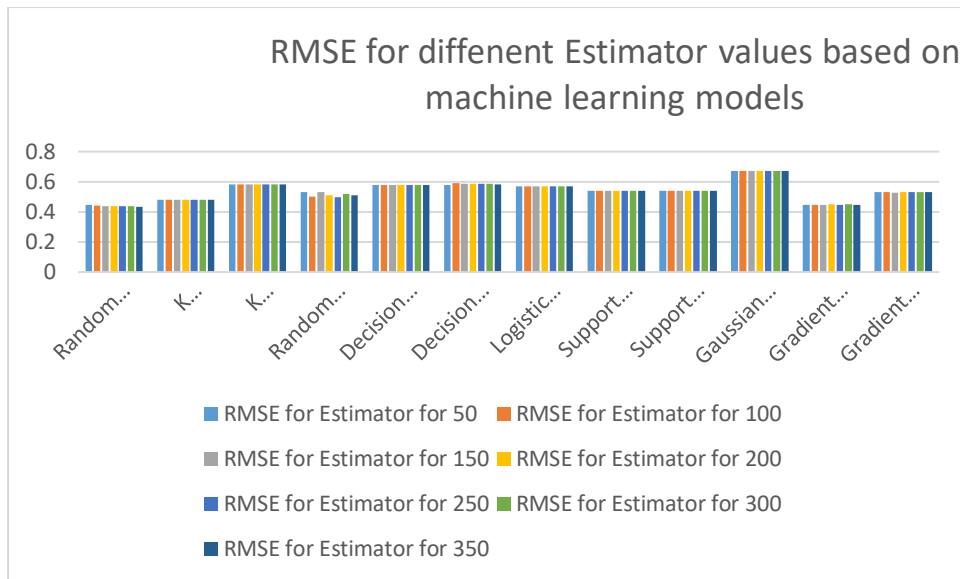
K Fold Cross validation using different estimator values.

A number of estimator values has been applied for a number of machine learning models. The root mean square value corresponding to machine learning models have been furnished in table 7.

Table 7**Machine Learning model wise RMSE value for a number of estimator values**

| Model | RMSE for Estimator for 50 | RMSE for Estimator for 100 | RMSE for Estimator for 150 | RMSE for Estimator for 200 | RMSE for Estimator for 250 | RMSE for Estimator for 300 | RMSE for Estimator for 350 |
|--|---------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Random Forest Regressor | 0.446 | 0.442 | 0.4390 | 0.437 | 0.4367 | 0.4377 | <u>0.435</u> |
| K Neighbors Regressor | 0.4789 | 0.4789 | 0.4789 | 0.4789 | 0.4789 | 0.4789 | <u>0.4789</u> |
| K Neighbors classifier | 0.582 | 0.582 | 0.582 | 0.582 | 0.582 | 0.582 | <u>0.582</u> |
| Random Forest Classifier | 0.5305 | 0.502 | 0.5304 | 0.5107 | <u>0.499</u> | 0.5192 | 0.5107 |
| Decision Tree Regressor | 0.5776 | 0.5776 | 0.5776 | 0.5776 | 0.5776 | 0.5776 | <u>0.5776</u> |
| Decision Tree Classifier | <u>0.577</u> | 0.5923 | 0.5853 | 0.5849 | 0.5851 | 0.5848 | 0.5801 |
| Logistic Regression | 0.5676 | 0.5676 | 0.5676 | 0.5676 | 0.5676 | 0.5676 | <u>0.5676</u> |
| Support Vector Machine(kernel = linear) | 0.5413 | 0.5413 | 0.5413 | 0.5413 | 0.5413 | 0.5413 | <u>0.5413</u> |
| Support Vector Machine(kernel = rbf) | 0.5413 | 0.5413 | 0.5413 | 0.5413 | 0.5413 | 0.5413 | <u>0.5413</u> |
| Gaussian NB | 0.6717 | 0.6717 | 0.6717 | 0.6717 | 0.6717 | 0.6717 | <u>0.6717</u> |
| Gradient Boosting Regressor | 0.4480 | 0.4478 | 0.4480 | 0.4485 | 0.4479 | 0.4490 | <u>0.4476</u> |
| Gradient Boosting Classifier | 0.5301 | 0.5301 | <u>0.5273</u> | 0.5327 | 0.5301 | 0.53 | 0.5328 |

From table 7, it has been observed that the root mean square values for estimator values as 350 is minimum for most of the cases. Out of all machine learning models random forest regressor produces minimum root mean square value. Therefore, the acceptable machine learning model is Random Forest Regressor. The change values of RMSE based on a number of estimated values using machine learning models has been furnished in graph 5.



Graph 5

Stratified K-fold cross Validation.

Using stratified K-fold, the accuracy values based on machine learning models have been furnished in table 8.

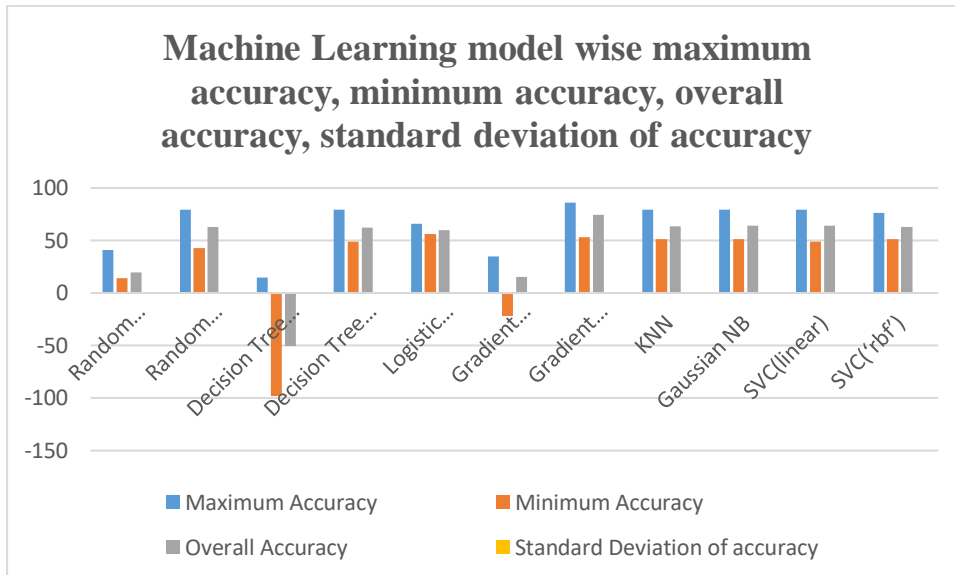
Table 8

Machine Learning model wise maximum accuracy, minimum accuracy, overall accuracy, standard deviation of accuracy

| Machine Learning Models | Maximum Accuracy | Minimum Accuracy | Overall Accuracy | Standard Deviation of accuracy |
|------------------------------|------------------|------------------|------------------|--------------------------------|
| Random Forest Regressor | 41.02 | 13.86 | 19.486 | 0.1717 |
| Random Forest Classifier | 79.4117 | 42.857 | 62.958 | 0.10965 |
| Decision Tree Regressor | 14.99 | -98.33 | -50.78 | 0.3845 |
| Decision Tree Classifier | 79.41 | 48.57 | 62.06 | 0.0999 |
| Logistic Regression | 65.71 | <u>55.88</u> | 59.71 | <u>0.0363</u> |
| Gradient Boosting Regressor | 35.036 | -21.7 | 15.326 | 0.1718 |
| Gradient Boosting Classifier | <u>85.71</u> | 52.94 | 74.15 | 0.097 |
| KNN | 79.4117 | 51.428 | 63.52 | 0.0946 |
| Gaussian NB | 79.4117 | 51.428 | <u>63.815</u> | 0.0995 |
| SVC(linear) | 79.4117 | 48.571 | 63.789 | 0.09831 |
| SVC('rbf') | 76.47 | 51.42 | <u>62.93</u> | 0.07739 |

From table 8, it has been observed that gradient boosting classifier is preferred on the basis of maximum accuracy, logistic regression on the basis of minimum accuracy, Gaussian NB on

the basis of average accuracy, logistic regression on the basis of standard deviation of accuracy. It can be mentioned that more positive values are preferred in case of maximum, minimum and average accuracy values and minimum value is preferred in case of standard deviation of accuracy. The change values of maximum accuracy, minimum accuracy, overall accuracy, standard deviation of accuracy have been furnished in graph 6.



Graph 6

Leave One Out Method.

Using leave one out, the accuracy values based on machine learning models have been furnished in table 9.

Table 9
Machine Learning model wise accuracy using Leave one out method

| Machine Learning model | Accuracy |
|--|----------|
| Linear Regression | 0.44 |
| Logistic Regression | 0.44 |
| Random Forest Classification | 0.25 |
| Random Forest Regressor | 0.38 |
| Decision Tree Classifier | 0.35 |
| Decision Tree Regressor | 0.34 |
| Gaussian NB | 0.45 |
| KNN Classifier | 0.34 |
| Support Vector Machine (Kernel = Linear) | 0.29 |
| Support Vector Machine (Kernel = RBF) | 0.40 |

From table 9, Gaussian NB has highest accuracy value of 0.40. Therefore, the preferable model is Gaussian NB.

Hold out Method.

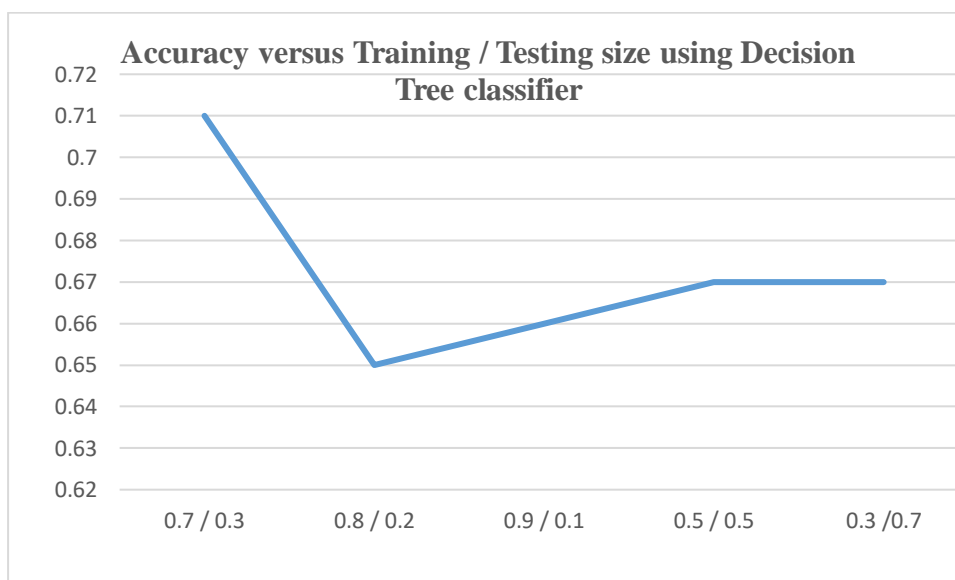
Initially decision Tree classifier has been applied using different training and test size and the corresponding accuracy has been furnished in table 10. In table 10, columns 1, 3, 5, 7, 9 stand

for the ratio of training / test size. Columns 2, 4, 6, 8, 10 stand for corresponding accuracy values.

Table 10
Accuracy versus Training / Testing size using Decision Tree classifier

| Train/Test | Accur. | Train /Test | Accur . | Trai n/Te st | Accur. | Trai n/T est | Accur. | Train /Test | Accur. |
|------------|-------------|-------------|---------|--------------|--------|--------------|--------|-------------|--------|
| 0.7/0.3 | <u>0.71</u> | 0.8/0.2 | 0.65 | 0.9/0.1 | 0.66 | 0.5/0.5 | 0.67 | 0.3/0.7 | 0.67 |

From table 10, it has been observed that maximum accuracy value is 0.71 corresponding training and test size as 70:30. Therefore this training and test size has to be used in other machine learning models. The change of value of accuracy based on training size / testing size using decision tree classifier has been furnished in graph 7.



Graph 7

The accuracy based on these models using hold out accuracy have been furnished in table 11.

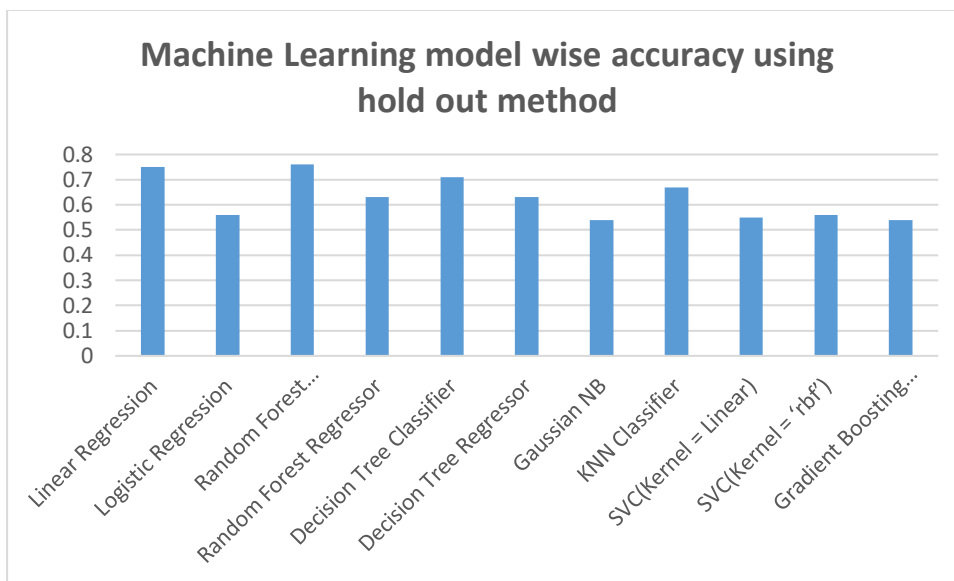
Table 11
Machine Learning model wise accuracy using hold out method

| Machine Learning model | Accuracy |
|------------------------------|-------------|
| Linear Regression | 0.75 |
| Logistic Regression | 0.56 |
| Random Forest Classification | 0.76 |
| Random Forest Regressor | 0.63 |
| Decision Tree Classifier | 0.71 |
| Decision Tree Regressor | 0.63 |
| Gaussian NB | 0.54 |
| KNN Classifier | 0.67 |

| | |
|-------------------------------------|-------------|
| SVC(Kernel = Linear) | 0.55 |
| SVC(Kernel = 'rbf') | 0.56 |
| Gradient Boosting Classifier | 0.54 |

On the basis of table 11, it has been observed that the accuracy based on Random Forest Classification algorithm is maximum, therefore the preferred model is Random Forest Classifier.

The training accuracy and testing accuracy based on ridge model and lasso model have been furnished in table 12. The change of value of accuracy based machine learning models has been furnished in graph 8.



Graph 8

Table 12
Training and Testing Accuracy based on Ridge and Lasso model

| Model | Training Accuracy | Testing Accuracy |
|-------------|-------------------|------------------|
| Ridge model | 0.18 | -0.02 |
| Lasso model | 0.0 | -0.0266 |

Recursive Feature Elimination.

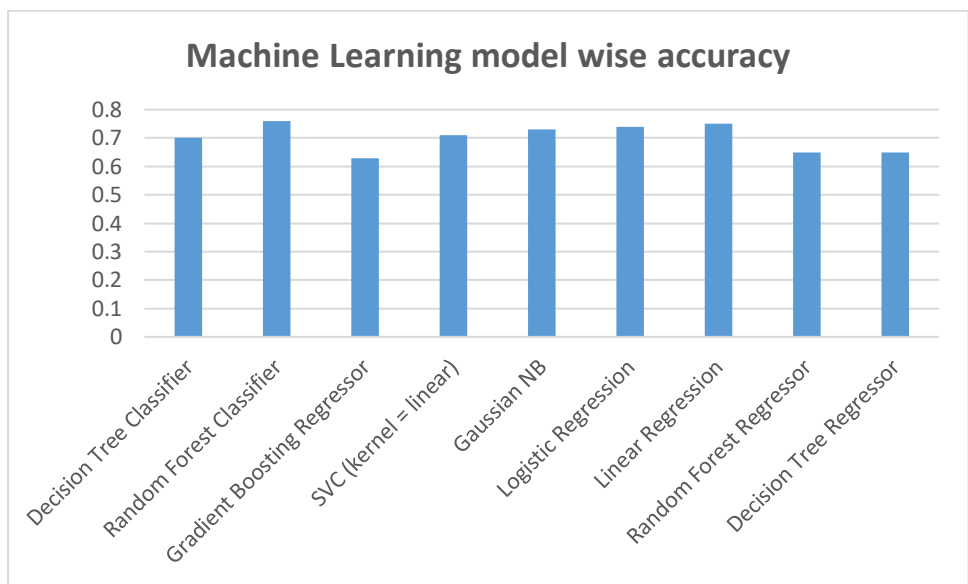
Recursive Feature Elimination is a feature selection method to identify a dataset's key features. The process involves developing a model with the remaining features after repeatedly removing the least significant parts until the desired number of features is obtained. Machine learning models have been applied using recursive feature elimination. The number of features, details of main features and accuracy have been furnished in table 13 corresponding to lever data set.

Table 13
Machine Learning model wise accuracy using Recursive Feature Elimination

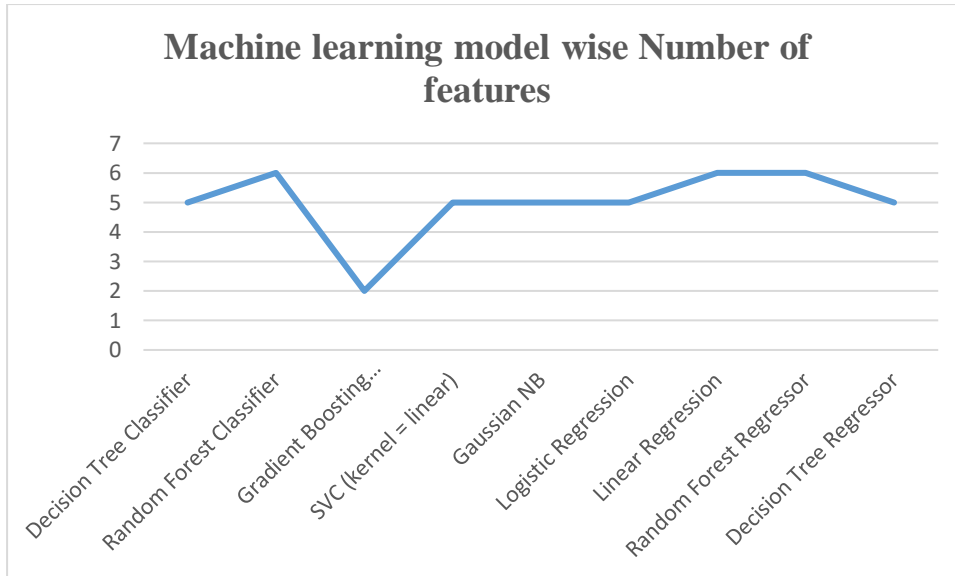
| Machine Learning Models | No of features | Main Features | Accuracy |
|--------------------------|----------------|---------------|-------------|
| Decision Tree Classifier | 5 | T T T T T F | 0.7 |
| Random Forest Classifier | 6 | T T T T T T | <u>0.76</u> |

| | | | |
|-----------------------------|---|-------------|------|
| Gradient Boosting Regressor | 2 | F F T F T F | 0.63 |
| SVC (kernel = linear) | 5 | T F T T T T | 0.71 |
| Gaussian NB | 5 | T F T T T T | 0.73 |
| Logistic Regression | 5 | T T T T F T | 0.74 |
| Linear Regression | 6 | T T T T T T | 0.75 |
| Random Forest Regressor | 6 | T T T T T T | 0.65 |
| Decision Tree Regressor | 5 | T T T T T F | 0.65 |

From table 13, it has been observed that Random Forest Classifier has maximum accuracy value as compared to other models. Therefore, the preferred model is Random Forest Classifier. The change of value of accuracy based on machine learning models using recursive feature elimination has been furnished in graph 9. The change of value of number of features based on machine learning models using recursive feature elimination has been furnished in graph 10.



Graph 9



Graph 10

8. Conclusion.

From table 3, it has been observed that performance of Support vector machine with linear kernel and linear regression model is excellent as compared to other models. The advantage of Support vector machine with linear kernel is that it can perform well at classifying non-linear data, it can reduce the overfitting of data, it can learn without a local minima, it can perform well on data sets that have many attributes.

The advantage of linear regression is simple Implementation, performs best on Linear Data, overfitting can be reduced by regularization. Comparing these the preference of Support vector machine with linear kernel is more than linear regression.

From table 6, it has been observed that the performance of random forest regressor is better as compared to other models on the basis of minimum root mean square error value. From table 7, the estimator value as 350 gives minimum root mean square error value for random forest regressor as compared to other models. Therefore, the preferred model is random forest regressor. From table 8, gradient boosting classifier shows the maximum accuracy value as compared to other models. Using leave one out method, from table 9, Gaussian NB gives maximum accuracy value. From table 11 using hold out accuracy method, random forest classifier shows maximum accuracy value as compared to other models. From table 13 using recursive feature elimination, random forest classifier shows maximum accuracy value as compared to other models.

Therefore, it has been observed that out of 6(six) cases, random forest classification has been preferred by 2 cases, random forest regression has also been preferred by 2 cases. Gaussian NB and gradient boosting classifier has been preferred by 1(one) case. Comparing with of random forest classifier, it has been found that random forest classifier is the best machine learning model among others. The advantage of random forest classifier is that it is robust to Over fitting, it can handle missing values, it can give feature Importance i.e. it can provide insights into those features which are most influential in predictions, it can help in aiding feature selection. Random Forest can handle large datasets with numerous features

and data points, making it versatile for various applications. Considering all these points, it is preferable to choose random forest classifier for prediction of liver disease.

The primary distinction between a Random Forest Classifier and a Random Forest Regressor lies in the nature of the prediction task they address: classification versus regression.

In Random Forest Classifier, the model is designed for classification tasks, where the goal is to predict categorical outcomes. It operates by aggregating the predictions of multiple decision trees through majority voting. Each tree in the forest contributes a class label, and the class receiving the most votes becomes the final prediction.

In Random Forest Regressor, this model is useful for regression tasks, aiming to predict continuous numerical values. It combines the predictions of individual decision trees by averaging their outputs, resulting in a continuous numerical prediction.

Both models utilize ensemble learning techniques to enhance predictive accuracy and robustness. While the classifier focuses on categorical outcomes with majority voting, the regressor emphasizes continuous outcomes through averaging.

References

- [1] T. K. Lakshmi, Dheeba J, “ Predictive Analysis of Periodontal Disease Progression Using Machine Learning: Enhancing Oral Health Assessment and Treatment Planning”, *International Journal of Intelligent Systems and Applications in Engineering*, IJISAE, 2023, 11(10s), 660–671, ISSN:2147-67992.
- [2] Dr.Avinash J. Agrawal, Dr.Rashmi R. Welekar, Dr.NamitaParati, Dr.Pravin R. Satav, Dr.Leena H. Patil, Dr.Shailendra S. Aote, “ Diabetes Prediction Using Medical Data and Disease Influence Measures using Machine Learning”, *International Journal of Intelligent Systems and Applications in Engineering*, IJISAE, 2023, 11(10s), 01–10, ISSN:2147-67992.
- [3] Dr.Araddhana A. Deshmukh, Dr.KirtiWanjale, Tushar A. Jadhav, Dr.Dhananjay V. Khankal⁴, Dr. Ajay D. Diwate⁵, Shashikant V. Athawale⁶, “ Multi-Class Skin Diseases Classification Using Hybrid Deep Convolutional Neural Network”, *International Journal of Intelligent Systems and Applications in Engineering*, ISSN:2147-67992147-6799 www.ijisae.org, IJISAE, 2023, 11(10s), 11–22.
- [4] VivekVeeraiah, Dr.Ravikaumar G K, Neeraj Gupta, Dr. Dinesh Singh, Mr. Vinod Motiram Rathod, Rama Krishna Yellapragada, “ Diagnosing of Disease using Machine Learning in Healthcare by Internet of Things”, *International Journal of Intelligent Systems and Applications in Engineering*, ISSN:2147-6799, ISSN:2147-6799 IJISAE, 2023, 11(10s), 954–973.
- [5] K. Mary Sudha Rani, Dr. V. Kamakshi Prasad, “ Identification of Lung Cancer Using Ensemble Methods Based on Gene Expression Data”, *International Journal of Intelligent Systems and Applications in Engineering*, ISSN:2147-67992147-679, IJISAE, 2023, 11(10s), 257–266.
- [6] VarshaBhole, Harsh NamdevBhor, Jayprabha Vishal Terdale, VijayaPinjarkar, RashmiMalvankar, Neha Zade, “ Machine Learning Approach for Intelligent and Sustainable Smart Healthcare in Cloud-Centric IoT”, *International Journal of Intelligent Systems and Applications in Engineering*, ISSN:2147-67992147- IJISAE, 2023, 11(10s), 36–48.
- [7] RupaliAtul Mahajan, Dr.BalasahebBalkhande, Dr.KirtiWanjale, Dr.AbhijitChitre, TusharAnkushJadhav, Dr. Sheela NarenHundekari, “Enhancing Heart Disease Risk Prediction Accuracy through Ensemble Classification Techniques”, *International Journal of Intelligent Systems and Applications in Engineering*, ISSN:2147-67992147-6799 IJISAE, 2023, 11(10s), 701–713.
- [8] Dr. H. E. Khodke, Dr. Manoj Bhalerao, Dr. S. N. Gunjal, Dr.SuvarnaNirmal, Santosh Gore, Dr. B. J. Dange, “ An Intelligent Approach to Empowering the Research of Biomedical Machine Learning in Medical Data Analysis using PALM”, *International Journal of Intelligent*

Systems and Applications in Engineering,ISSN:2147-67992 , IJISAE, 2023, 11(10s), 429–436.

[9] B. Reuben and C. Narmadha, “ Effective Kidney Stone Prediction Based on Optimized Yolov7 Segmentation and Deep Learning Classification”, International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-67992 IJISAE, 2024, 12(1), 183–192.

[10] LeenaChaudhari, Dr.Dhananjay E. Upasani, Swati Gawhale, Dr.Dhananjay V. Khankal, Jambi Ratna Raja Kumar, Nishant A. Upadhye, “Analysis of Critical Diseases from ECG Signal Using Hybrid CNN and LSTM”,International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-67992147-6799 IJISAE, 2023, 11(10s), 217–232.

[11] S.Gopalakrishnan, M. Sahaya Sheela , K. Saranya, J. Jasmine Hephzipah, “ A Novel Deep Learning-Based Heart Disease Prediction System Using Convolutional Neural Networks (CNN) Algorithm”, International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-6799, IJISAE, 2023, 11(10s), 516–522.

[12] SaruchiKukkar, Japreet Singh, “ Breast Cancer Detection and Classification by Features Non-Linear Mapping with Random Forest Classifier”, , International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-6799, IJISAE, 2024, 12(1), 193–202.

[13] SaruchiKukkar, Japreet Singh, “Breast Cancer Detection and Classification by Features Non-Linear Mapping with Random Forest Classifier”,International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-6799 2147-6799 IJISAE, 2024, 12(1), 193–202.

[14] Ms. Seema B. Rathod, Dr.Lata L. Ragha, “The Detection of Lung Tumors Using CT scan Images with Feature Extraction and Segmentation Techniques.”,International Journal of Intelligent Systems and Applications in Engineering ISSN:2147-6799214, IJISAE, 2024, 12(1), 628–638.

[15] Dr.PallaviHallappanavarBasavaraja, Dr.NandeeswarSampigehalliBasavaraju ,Dr. Pooja Nayak S., Dr.AnushaPreetham, Dr.Ramya R. S.,Prof. Shravya S. “A Framework for Brain Tumor Image Analysis using Convolution with RELU”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-67992, IJISAE, 2024, 12(3), 312–321.

[16] AbdulrahmanArishi , Suma Alex Kanjramnilkunathil , Sudha K. Rajan, AfshanKausar, ArshiaArjumand and Fred Torres-Cruz, “Cardiac Abnormalities Classification Model Using Improved Deep Learning Approach”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(1), 360–367.

[17] Srikanth R, Tamil Priya D., Jagadeesan S., Savita P. Patil, Anupama K. Ingale, ManojkumarVivekanandan, VenkadeshRamalingam, “A Comprehensive Review on Cancer Prediction Using Machine Learning Techniques”,International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(3), 115–127.

[18] Archana J. Jadhav, Dipali. H. Patil, G. S. Mate, R. A. Deshmukh, Anjali S. More, Chandan Prasad, “Detection of Brain Tumor using Fine-Tuned Pre-Trained MobileNet Deep Learning Model”, International Journal of Intelligent Systems and Applications in EngineeringISSN:2147-6799214, IJISAE, 2024, 12(3), 361–368.

[19] Aditya Gupta, Angad Singh, MaharshiJani, YuvrajSalaria, Dr. Vani Hiremani, Dr.SudhanshuGonge, Dr.KetanKotecha, “Diabetes Prediction and Apprehension with Focus Both on Clinical and Non-Clinical Factors”,International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(1), 746–755.

[20] V. Kavitha, K. Ulagapriya, “Comparative Evaluation for Brain Tumor Detection Using Inception-V3 Architecture”, International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-6799214, IJISAE, 2024, 12(1), 277–283.

[21] SujeethBabuKolli, Vinay Kalisetti, Rakesh Varaparla, Vamsi Sai Chandra Vasarla, VeeraswamyAmmisetty, “Automated Diagnosis of Diabetic Retinopathy using Deep Learning

and Image Analysis”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(3), 174–179.

[22] Hajer Sayed Hussein, Hussein AlBazar, Roxane Elias Mallouhy, Fatima Al-Hebshi, “Deep Learning in Heart Murmur Detection: Analyzing the Potential of FCNN vs. Traditional Machine Learning Models”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1296-1304

[23] Mohammed A M Abueed¹, DanialMd Nor², Nabilah Ibrahim³, Jean-Marc Ogier⁴, “Pneumonia Detection Using Transfer Learning: A Systematic Literature Review”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1032-1041.

[24] JanjhyamVenkata Naga Ramesh, P N S Lakshmi, Dr.ThalakolaSyamsundararao, ElangovanMuniyandy, LinginediUshasree, Prof.Ts. Dr. Yousef A.Baker El-Ebiary, Dr. David NeelsPonkumarDevadhas, “Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1011-1023.

[25] NandakumarPandiyam, Subhashini Narayan, “Comparative Analysis of Cardiac Disease Classification Using a Deep Learning Model Embedded with a Bio-Inspired Algorithm”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 976-986.

[26] Muhammad Sajid, Wareesa Sharif, Ghulam Gilanie, Maryam Mazher, Khurshid Iqbal, Muhammad Afzaal Akhtar, Muhammad Muddassar, Abdul Rehman, “IoMT-Enabled Noninvasive Lungs Disease Detection and Classification Using Deep Learning-Based Analysis of Lungs Sounds”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 877-886.

[27] LastrWidyaAstuti, Ermatita, Dian PalupiRini, “Handling Imbalanced Data in Medical Records Using Entropy with Minkowski Distance”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025 , 867-876.

[28] U. Sudha Rani, Dr. C. Subhas, “Enhanced Early Detection of Diabetic Nephropathy Using a Hybrid Autoencoder-LSTM Model for Clinical Prediction”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 840-849.

[29] JanjhyamVenkata Naga Ramesh, F. Sheeja Mary, Dr. S. Balaji, Dr.DivyaNimma, ElangovanMuniyandy, A.SmithaKranthi, Prof.Ts. Dr. Yousef A.Baker El-Ebiary, “Enhanced Colon Cancer Prediction Using Capsule Networks and Autoencoder-Based Feature Selection in Histopathological Images”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 817-829.

[30] Marwen SAKLI, Chaker ESSID, Bassem BEN SALAH, Hedi SAKLI, “Flexible Framework for Lung and Colon Cancer Automated Analysis Across Multiple Diagnosis Scenarios”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 569-580.

[31] Noha Ayman, Mahmoud E. A. Gadallah, Mary MonirSaeid, “Multi-Classification Convolution Neural Network Models for Chest Disease Classification”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 373-380.

[32] Shi Li, Xueyao Sun, “Dialogue-Based Disease Diagnosis Using Hierarchical Reinforcement Learning with Multi-Expert Feedback”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 298-306.

[33] Wilmer Alberto Pacheco Llacho, Eveling Castro-Gutierrez, Luis David Huallpa Tapia, “Lung Parenchyma Segmentation Using Mask R-CNN in COVID-19 Chest CT Scans”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 141-146.

- [34] Manoj Yadav Mamilla, Ronak Al-Haddad, Stiphen Chowdhury, “Resampling Imbalanced Healthcare Data for Predictive Modelling”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 36-44.
- [35] Anshul Jain, Vikas Sakalle, “A Review of Glaucoma Optic Disk Localization and Classification Machine Learning and Deep Learning Models”, International Journal of Innovative Research in Technology and Science, ISSN: 2321-1156 www.ijirts.org Volume 12 Issue 2, March 2024, 163-176.
- [36] Anchal Rajak, Nitish Kumar Roy, “Enhancing Machine Learning Through Neural Networks: A Comprehensive Exploration”, International Journal of Innovative Research in Technology and Science, ISSN: 2321-1156 www.ijirts.org Volume 12 Issue 2, March 2024, 155-162.
- [37] Bhoopendra Singh, Subodhini Gupta, “Prediction of Cardiovascular Disease Using Machine Learning”, International Journal of Innovative Research in Technology and Science, ISSN: 2321-1156 www.ijirts.org Volume 12 Issue 2, March 2024, 1-7.
- [38] Baharuddin, S Gultom, M Restuati, A Mansyur, H Fibriasari, I I Pane, S Pujiastuti, H D Hutahaean, B D Waluyo, V S G Galamgam, R PHacla, M A Rais, R Effendi, M R Pane, “Implementation of a convolutional neural network method to predict university students’ emotional extent in online learning”, International Journal of Engineering & Technology, 11 (1) (2022) 59-62.
- [39] Signe Feudjeu Josias Éric, Nzié Wolfgang, Ngnassi Djami Aslain Brisco, Ntsama Eloundou Pascal, Chassem Priva, “Modelling of improved LSTM +1D convolution neural network methods for the diagnosis of SKF bearings”, International Journal of Engineering & Technology, 13 (2) (2024) 194-203.
- [40] Anil Kumar Prajapati, Umesh Kumar Singh, Rekha Singh, Arpita Shukla, “Study and analysis of feature selection problems and impact of bias in machine learning disease prediction models”, International Journal of Engineering & Technology, 13 (1) (2024) 182-188.
- [41] Alwatben Batoul Rashed A, Abeer Alhujaylan, “A comparative study of multi-objective pso-fuzzy optimization with weka classification algorithms to improve the interpretability and accuracy in medical data”, International Journal of Engineering & Technology, 13 (1) (2024) 36-41.
- [42] Sangeetha Balasubramanian, Periyasamy Rajadurai, “Machine Learning-Based Classification of Pulmonary Diseases through Real-Time Lung Sounds”, International Journal of Engineering and Technology Innovation, vol. 14, no. 1, 2024, pp. 85-102.
- [43] Ajay Kumar, Kamaldeep Kaur, “A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction”, International Journal of Engineering and Technology Innovation, vol. 14, no. 1, 2024, pp. 29-43.
- [44] Sheena Christabel Pravin, Sindhu Priya Kanaga Sabapathy, Suganthi Selvakumar, Saranya Jayaraman, Selvakumar Varadharajan Subramani, “An Efficient DenseNet for Diabetic Retinopathy Screening”, International Journal of Engineering and Technology Innovation, vol. 13, no. 2, 2023, pp. 125-136.
- [45] Chuan-Pin Lu, Yan-Long Huang, Po-Jen Lai, “Development of the Abnormal Tension Pattern Recognition Module for Twisted Yarn Based on Deep Learning Edge Computing”, International Journal of Engineering and Technology Innovation, vol. 13, no. 4, 2023, pp. 284-295.
- [46] Laize D.L. Trindade; Diovana G. de Batista; Maira S. Brigo; Matias N. Frizzo; Rafael Z. Frantz; Fabricia Roos-Frantz; Thiago G. Heck; Sandro Sawicki, “Machine learning applied in blood laboratory database for identification of an obesogenic/diabetogenic diet consumption: a preclinical modelling approach”, International Journal of Computer Applications in Technology (IJCAT), Vol. 75, No. 1, 2024.

- [47] Ahmad Taher Azar, “A bio-inspired method for segmenting the optic disc and macula in retinal images”, International Journal of Computer Applications in Technology (IJCAT), Vol. 72, No. 4, 2023.
- [48] S. Pratyaksha; M.L. Chayadevi, “Diagnosis and management of arrhythmia using machine learning”, International Journal of Computer Applications in Technology (IJCAT), Vol. 72, No. 2, 2023.
- [49] Musa Alshawabkeh, Mohammad Hashem Ryalat, Osama M. Dorgham, Khalid Alkharabsheh, Mohammad HjoujBtoush, MamounAlazab, “A hybrid convolutional neural network model for detection of diabetic retinopathy”, <https://doi.org/10.1504/IJCAT.2022.130886>.
- [50] Ambika Annavarapu, Surekha Borra, “Figure-ground segmentation based medical image denoising using deep convolutional neural networks”, <https://doi.org/10.1080/1206212X.2024.2420870>.
- [51] RishavPramanik , Momojit Biswas , Shibaprasad Sen , Luis Antonio de Souza Júnior , João Paulo Papa , Ram Sarkar , “A fuzzy distance-based ensemble of deep models for cervical cancer detection”, Computer Methods and Programs in Biomedicine, Computer Methods and Programs in Biomedicine 219 (2022) 106776, www.elsevier.com/locate/cmpb.
- [52] Ruhul Amin and Rubia Yasmin and Sabba Ruhi and Md Habibur Rahman and Md ShamimReza,"Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning approaches" Informatics in Medicine Unlocked, <https://doi.org/10.1016/j.imu.2022.101155>
Medicine Unlocked, <https://doi.org/10.1016/j.imu.2023.101369> url: <http://www.ijpam.eu>.
- [53] Madhab Paul Choudhury, J. Paul Choudhury, “A Framework for Developing Correct Software Programs through Software Defect Prediction and Elimination using Machine Learning Models”, International Journal for Research in Applied Science Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 12 Issue IX Sep 2024, page 402-411
- [54] Madhab Paul Choudhury, J. Paul Choudhury, “A comparative study on the performance of Soft Computing models in the prediction of Orthopedic disease in the environment of Internet of Things” . Proceedings of ICCM 2022 Department of Electronics and Communication Engineering and the Department of Electrical Engineering, North Eastern Regional Institute of Science and Technology (NERIST), Arunachal Pradesh, India in collaboration with Emlyon Business School France, July 2022 Applications of Computational Intelligence in Management Mathematics Page247-258. <https://link.springer.com/chapter/10.1007/978-3-031-25194-820>
- [55] 4 Ways to Evaluate your Machine Learning Model: Cross-Validation Techniques <https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/>
- [56] Liver Data Set: <https://www.kaggle.com/datasets/uciml/indian-liver-patient-records>