Urinary System Diseases Diagnosis Using Machine Learning Techniques
Term Paper- Group 6
University of Alabama in Huntsville

Introduction

The urinary system is the organ system that regulates fluid production, storage, and excretion. These systems include the kidneys, bladder and urethra. It represents the primary filtering system of the blood and any imbalance of this organ can lead to an increased risk of disease outbreaks. This paper presents a comprehensive review on the application of machine learning techniques for the diagnosis of urinary tract infections. The study focuses on the implementation of Support Vector Machine (SVM) and Artificial Neural Network (ANN) on the Acute Inflammations dataset obtained from the UCI Machine Learning Repository. The classification includes two diseases involving the urinary tract infection: inflammation and nephritis of the renal pelvic base.

The aim of the study is to evaluate the performance of these algorithms in terms of accuracy, precision, recall, and f1 score. Support vector machines use a different approach to segmentation, while artificial neural networks use the power of connected nodes to capture complex patterns in data. Through rigorous experimentation and cross-validation, the study attempts to identify the strengths and limitations of each method. This research not only furthers our understanding of the capabilities of SVM and ANN but also provides clinicians and researchers with valuable information to aid in the early and accurate detection of urinary disease.

Dataset

The Acute Inflammation dataset from the UCI Machine Learning Repository plays an important role in our research. This database includes the clinical information needed to diagnose a urinary tract infection (UTI) and consists of a total of 120 examples of importance to the disease. The attributes include temperature, nausea occurrence, lumbar pain presence, urine pushing, micturition pains, and the presence of blood in urine.

Each occurrence in the data set plays an important role in facilitating diagnosis, providing valuable insight into the patient’s condition. By using these factors, health care providers can make informed decisions about the incidence and severity of acute infection in UTI cases. The dataset's binary classification framework enables the categorization of patients into two classes: acute nephritis and acute inflammation, allowing for targeted and effective treatment strategies.

Related Work

Paper entitled "Prediction of Urinary System Disease Diagnosis: A Comparative Study of Three Decision Tree Algorithms" authored by Mahmoud Hussein Kadhem and Ahmed M. Kadhem of the University of Bahrain focuses on the use of data mining (DM) in developing a predictive model for the diagnosis of acute urologic cystitis and nephritis of the entire renal pelvis. The study examines three supervised machine learning algorithms. The dataset included in this study has several features that are crucial for the diagnosis of nephritis or acute bladder inflammation in any patient. To find the best classification algorithm that will be utilized to create an accurate prediction model, this study compares the performance and accuracy of the supervised machine learning algorithms Ridor, OneR, and J48. The decision tree (J48) has been utilized to categorize patient data into the appropriate acute inflammatory disorders due to its strong predictive accuracy and capabilities. The 10-fold cross validation method has

been used to train the examined dataset. A decision tree has been created for both nephritis and acute bladder infection.

The study also compares the effectiveness and precision of the supervised classification method, emphasizing the J48 decision tree technique's usefulness for developing predictive models. The accuracy, precision, timeliness of models, and sensitivity of several methods are compared. The research findings indicate that the J48 decision tree algorithm demonstrates remarkable accuracy and predictive capacity in the context of urinary tract infections. It achieves 100% precision and accuracy in the model prediction of acute urinary tract infections and pneumonia, among other acute diseases, thereby directing its focus towards the development of predictive models [1].

The paper titled "Urinary System Disease Diagnosis Using Machine Learning Techniques" authored by Seyyid Ahmed Medjahed,Tamazouzt Ait Saadiand Abdelkader Benyettou focuses on the implementation of machine learning algorithms for the diagnosis of the urinary system diseases. The aim of this paper is to evaluate the performance of different variants of Support Vector Machines and k-Nearest Neighbor with different distances and try to achieve a satisfactory rate of diagnosis. The two particular disorders under investigation are renal pelvic base nephritis and inflammation of the urinary bladder. This study is conducted using the "Acute Inflammations Data Set" from the UCI Machine Learning Repository. The effectiveness of the classification model is assessed using evaluation metrics that include classification accuracy rate, classification time, sensitivity, specificity, positive predictive values, and negative predictive values.

The paper looks into the computational architecture of several Support Vector Machines (SVMs), such as Least Squares SVM (LS-SVM) and Sequential Least Optimization (SVM-SMO).It also discusses the k-NN algorithm, and the process for diagnosing urinary infections. The findings of the experiment demonstrate that SVM-QP and SVM-SMO reached an astounding 100% classification accuracy rate. Despite their superior performance in terms of classification time, k-NN algorithms with varying distances also produced higher classification accuracy. In conclusion, the study contributes to the advancement of medical diagnostic methods and highlights the potential of ML algorithms in healthcare application [2].

The paper entitled "Detection of Acute Inflammation of Urinary Bladder and Acute Nephritis of Renal Pelvis Origin Using Artificial Neural Network" authored by Amina Aleta, Amra Džuho, and Faris Hrvat explains the creation of an artificial neural network (ANN) based expert system for the diagnosis of acute renal pelvis origin nephritis and bladder inflammation. Urinary tract illnesses frequently exhibit overlapping symptoms, making proper diagnosis and treatment challenging and time-consuming. The goal is to deliver a more effective and precise diagnosis based on subjective parameters supplied by the patient and other diagnostic parameters examined by the doctor by utilizing an expert system that makes use of artificial intelligence techniques. This study is conducted using the "Acute Inflammations Data Set" from the UCI Machine Learning Repository. The dataset consists of 120 samples which were used to train the algorithm to identify if a patient has both illnesses, just one disease, or is in good condition. The feed forward neural network produced an accuracy of 95.83%, a sensitivity of 94.44%, and a specificity of 100%, according to the results. The technology enables medical professionals to enter symptoms that patients have seen or reported in order to precisely diagnose patients and avoid making a mistaken diagnosis.

The significance of this ANN lies in terms of improving the amount of time spent by physicians and patients during diagnosis. The future work involve improving the system's ability even further by

including new disorders that raise a patient's risk of developing urinary tract infections and including more symptoms associated with these illnesses was also highlighted. In conclusion, the created artificial neural network (ANN) shows potential in delivering more precise diagnoses and avoiding misdiagnosis, improving the standard of treatment for patients with urinary tract disorders [3].

The research article entitled "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease" authored by Hira Khalid, Ajab Khan ,Muhammad Zahid Khan ,Gulzar Mehmood ,and Muhammad Shuaib Qureshi explains how to use a hybrid machine learning model to predict chronic kidney disease (CKD). The paper discusses the rising incidence of CKD and the demand for precise machine learning-based diagnosis approaches. To effectively predict CKD, the authors suggest a hybrid model that integrates gradient boosting, decision tree classifier, random forest classifier, and Gaussian Naive Bayes. The study's objectives are to determine the most accurate classifier and assess the best machine learning classification methods. The authors employ a tabular method to give a thorough analytical evaluation of relevant work and critically examine and rank current machine learning classification algorithms according to accuracy.

The designed architecture outlines how the suggested hybrid model is implemented using Python and necessary modules. To find the ideal feature combination for prediction, they employ the Pearson correlation feature selection approach on the UCI CKD dataset. The model uses random forest as the meta-classifier and includes base classifiers like gradient boosting, decision trees, and Gaussian Naive Bayes. In order to improve the accuracy and performance of the model, the authors stress the significance of cross-validation. The hybrid model outperformed all other classifier with 100% model accuracy. In summary, the research offers thorough understandings of how machine learning methods might be applied to precisely predict and diagnose chronic kidney disease, therefore addressing the pressing need for early identification and efficient intervention [4].

The research article "Predicting Chronic Kidney Disease Using Machine Learning Techniques" by Dibaba Adeba Debal and Tilahun Melak Sitote examines the use of machine learning techniques to predict chronic kidney disease (CKD), and highlights its importance and need to diagnose chronic kidney disease at early stages and also be accurate to reduce health complications and economic burden. The study focuses on binary and multiclass classifiers for step prediction, using machine learning models such as random forest (RF), support vector machine (SVM), decision tree (DT) to determine the most suitable prediction features for model training Feature selection methods including Univariate Feature Selection (UFS) and Iterative Feature Resolution with Cross-Validation (RFECV). The evaluation of the samples was subjected to tenfold cross-validation and was conducted using various performance metrics, such as accuracy, precision, recall, F1-score, sensitivity, and specificity.

The study showed that RFECV based on RF model exhibited high performance, achieving 99.8% accuracy for binary classification of CKD Furthermore, the study examined multiclass classification models, where RF and RFECV gave 79% accuracy for five class data sets Serum creatinine, blood urea nitrogen, hemoglobin, specific gravity etc. In conclusion, the study highlights the potential of these methods to help physicians accurately diagnose and predict disease stages, and provides insightful information on the use of machine learning for kidney targeting early prognosis of chronic disease [5].

Methodology

1. Data acquisition and preprocessing

The code retrieves the Acute Inflammation dataset from the UCI machine learning repository and preprocesses it. Categorical features are mapped to numeric values, and the data set is divided into features and targets.

- **Loading the dataset**

The dataset is retrieved from the UCI Machine Learning Repository using the fetch_ucirepo function.
The dataset consists of items and values, where items represent the independent variable and values represent the dependent variable(s) for prediction.

- **Data Pre-processing**

The data set must go through a data preprocessing step because it may contain missing values, duplicate data, or noisy data. They have a significant impact on our output if left uncorrected, so the data preprocessing step ensures that missing values and noisy data are reduced, ensuring that the data is quality and consistent. In preprocessing we use segmentation, standardization, data transformation, and data visualization techniques to preprocess the dataset.

- **Data transformation**

We used data transformation techniques to ensure compatibility with machine learning algorithms. It involved transforming categorical data statistically to allow for meaningful analyzes and models. We convert categorical features to numeric. For example, categorical features such as 'nausea', 'stiffness', 'urine pushing' are mapped to binary values (0 or 1) representing 'no' or 'yes' conditions.

- **Splitting**

The classification of the dataset is important for an unbiased assessment of predictive performance. Our dataset was divided into two parts of 80% training and 20% testing.

- **Standardization**

Scaling features are an important part of the data set modeling process. In this case, standardization emerges as the dominant method, reorganizing statistical distributions into standardized values. This helps to ensure equivalence of scales across the various items, promoting meaningful and comparable analysis. To standardize our data, we imported ‘StandardScaler’ from the ‘sklearn.preprocessing’ module of the sklearn library. This group of utilities is designed to standardize a dataset by converting it to a standardized format. The ‘fit_transform ()’ method, in the ‘StandardScaler’ class, made this process easier.

- **Data visualization**

To visualize the data, we used seaborn and the matplotlib.pyplot library.‘matplotlib.pyplot’ to create heatmap and histograms to visualize the distribution of each feature, and provide insight into their spread and skewness. Additionally, we used the 'ocean generation' library to create interactive heat maps, allowing us to explore relationships between features and identify potential relationships. These visualization techniques provided a detailed overview of the dataset to aid feature identification and understanding before analysis from model training.

2. Model Training

Support Vector Machine (SVM)

Support Vector Machine is a powerful supervised machine learning algorithm designed for both classification and regression tasks. SVM works by finding the optimal hyper plane that best separates different classes in the feature space. In the context of urinary tract infection diagnosis, One-vs-Rest classification algorithm is implemented. The One-vs-Rest Support Vector Machine (SVM) algorithm is a robust supervised learning method used for binary classification tasks, such as acute inflammation detection In our implementation, we trained separate SVM models for ranking inflammation and nephritis prediction using the sklearn SVM library project. Initially, the default parameter setting (C=1, kernel=radial basis function, gamma=1/n_features*X.var ()) resulted in over fitting, resulting in limited predictive power to address this, we used GridSearchCV to systematically find multiple parameter combinations, finally the best order (C=1, gamma=1) that yielded improved model performance. The advantages of the linear kernel function proved to be effective in capturing relevant patterns in the data, and the ability of SVM to detect discrimination accuracy between benign and malignant samples was enhanced. Overall, the ability of SVM to handle binary classification tasks highlights its usefulness in medical diagnostic applications such as acute inflammation detection.

Artificial Neural Network (ANN)

An Artificial Neural Network is a computational model inspired by the structure and functioning of the human brain. We implemented the ANN using the MLPClassifier from scikit-learn which is a powerful tool for classification tasks. The MLPClassifier allows a simple description of the network structure including the hidden layer and the number of nodes. The number of layers, neurons, and activation functions are adjusted during the training phase to optimize the model's ability to discern between malignant and benign cases. For our implementation, we used a neural network with two hidden layers with 64 and 32 nodes respectively. The MLP model achieved an impressive 100% accuracy for both inflammation and nephritis prediction tasks, demonstrating the effectiveness of ANN in medical classification tasks. The adaptation of the MLPClassifier allowed us to explore different network configurations and fine-tune the model for better performance.

3. Evaluation and Results

A variety of performance indicators, including accuracy, confusion matrices, and classification reports, are used to assess the trained models. To evaluate model performance with different training dataset sizes, learning curves are plotted. This helps in identifying problems like over fitting or under fitting and in understanding how model performance varies with increased training data sets. To simplify the dataset for visualization, Principal Component Analysis (PCA) is used to decrease the dataset's dimensionality to two dimensions. Scatter plots are then used to show the modified data so that the separation of classes according to the model predictions can be seen.

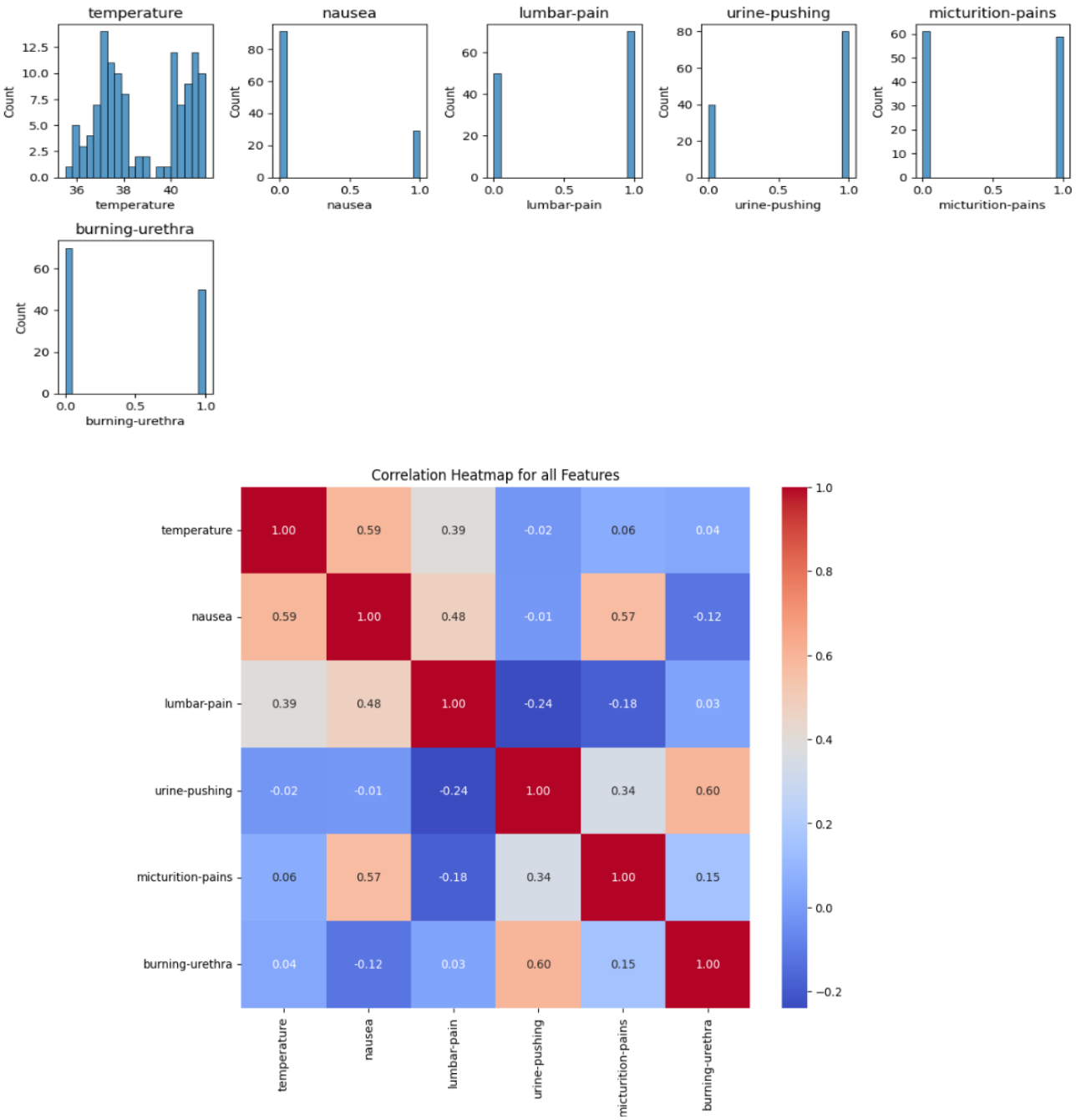
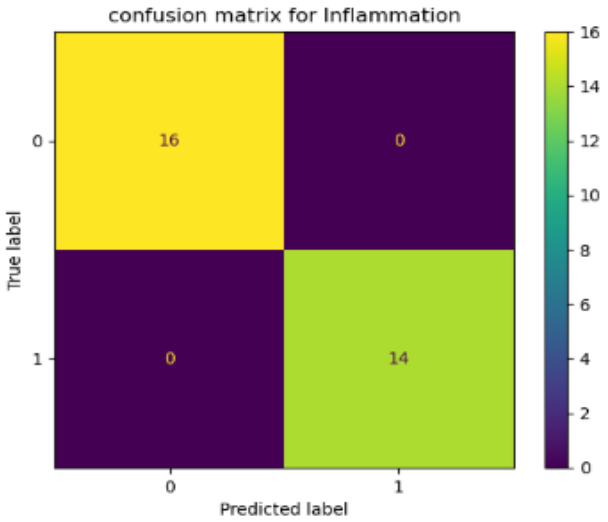
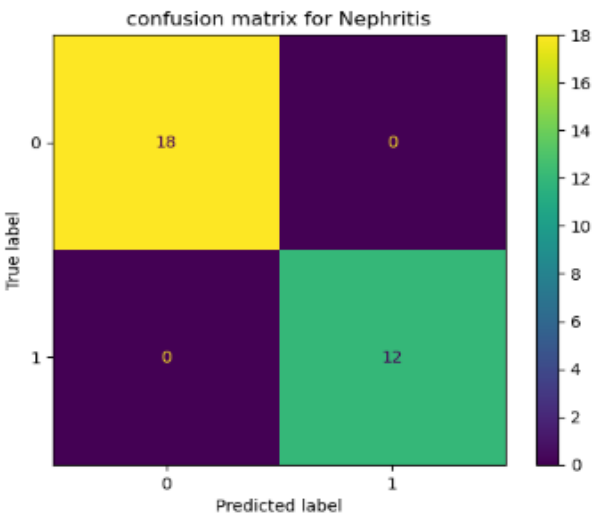
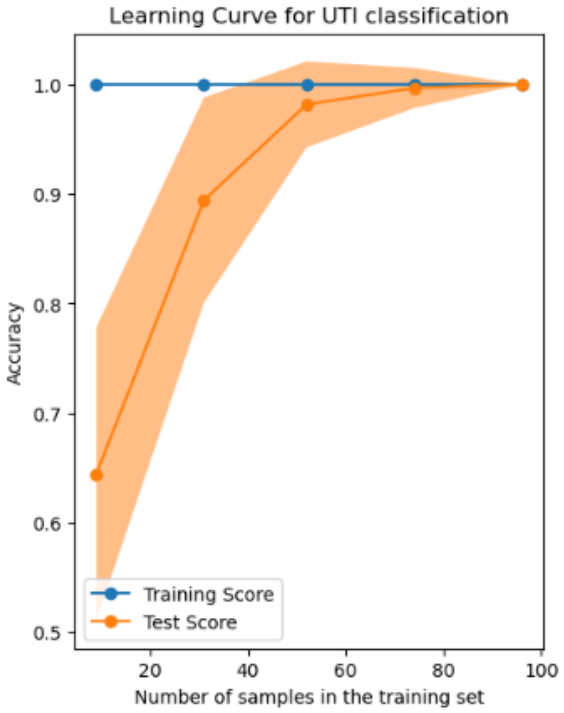


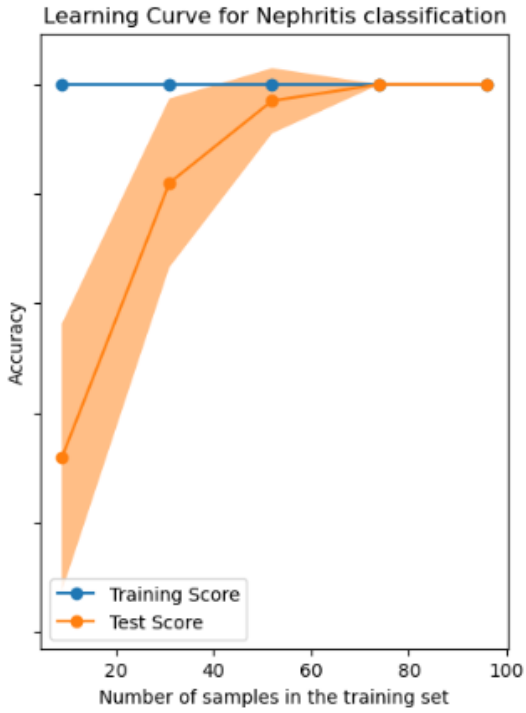
Fig 1: Data visualization

Fig 1 represents the data visualization of the acute inflammation dataset. Histograms and heatmaps are used to gain insights into the data distribution and relationships between variables. Heatmaps are used to visualize the correlation between features, helping to identify redundant or highly correlated features. Histograms are plotted for each feature to visualize their distributions and detect any anomalies.


confusion matrix for Inflammation
16
0
0
14
True label
Predicted label
0
1
0
1
0
2
4
6
8
10
12
14
16


confusion matrix for Nephritis
18
0
0
12
True label
Predicted label
0
1
0
1
0
2
4
6
8
10
12
14
16
18


Learning Curve for UTI classification
Accuracy
Number of samples in the training set
Training Score
Test Score
0.5
0.6
0.7
0.8
0.9
1.0
20
40
60
80
100


Learning Curve for Nephritis classification
Accuracy
Number of samples in the training set
Training Score
Test Score
20
40
60
80
100

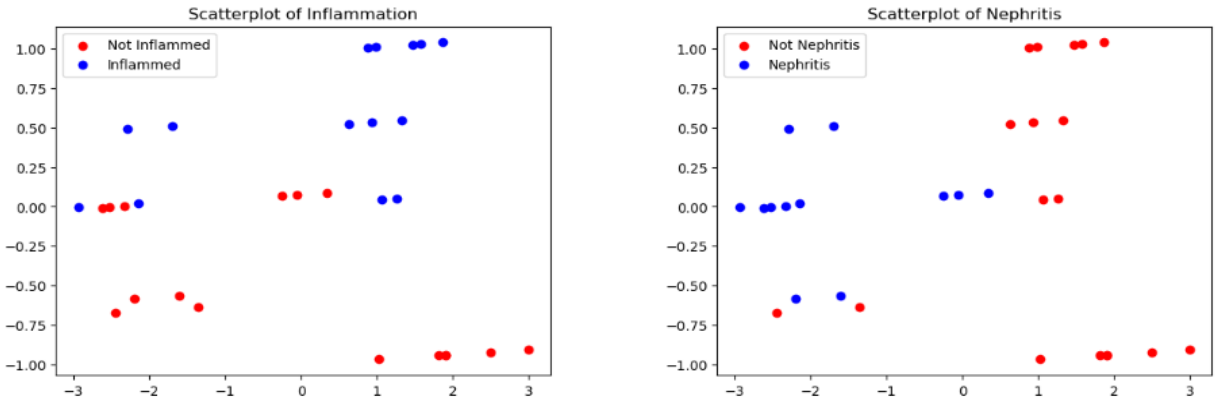
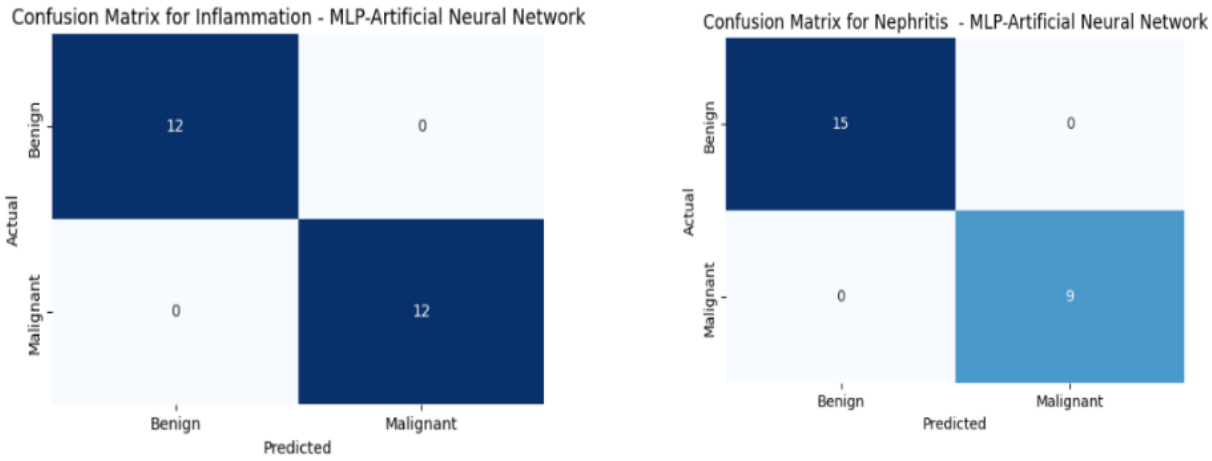


Fig 2: Results of SVM

Fig 2 represents the results of the SVM, the SVM model, after tuning hyper parameters using GridSearchCV; it achieved commendable accuracy (100%) and precision in distinguishing between cases of inflammation and nephritis. The confusion matrix visualizes insights into the model's ability to correctly classify instances into their respective classes, highlighting its robustness in handling both positive and negative cases. Furthermore, learning curves provided a visualization of the models' learning behavior, showcasing their ability to generalize well to unseen data and highlighting any potential over fitting or under fitting tendencies. Scatter plots further illustrated the distribution of predicted classes, offering insights into the models' classification performance in a two-dimensional space.



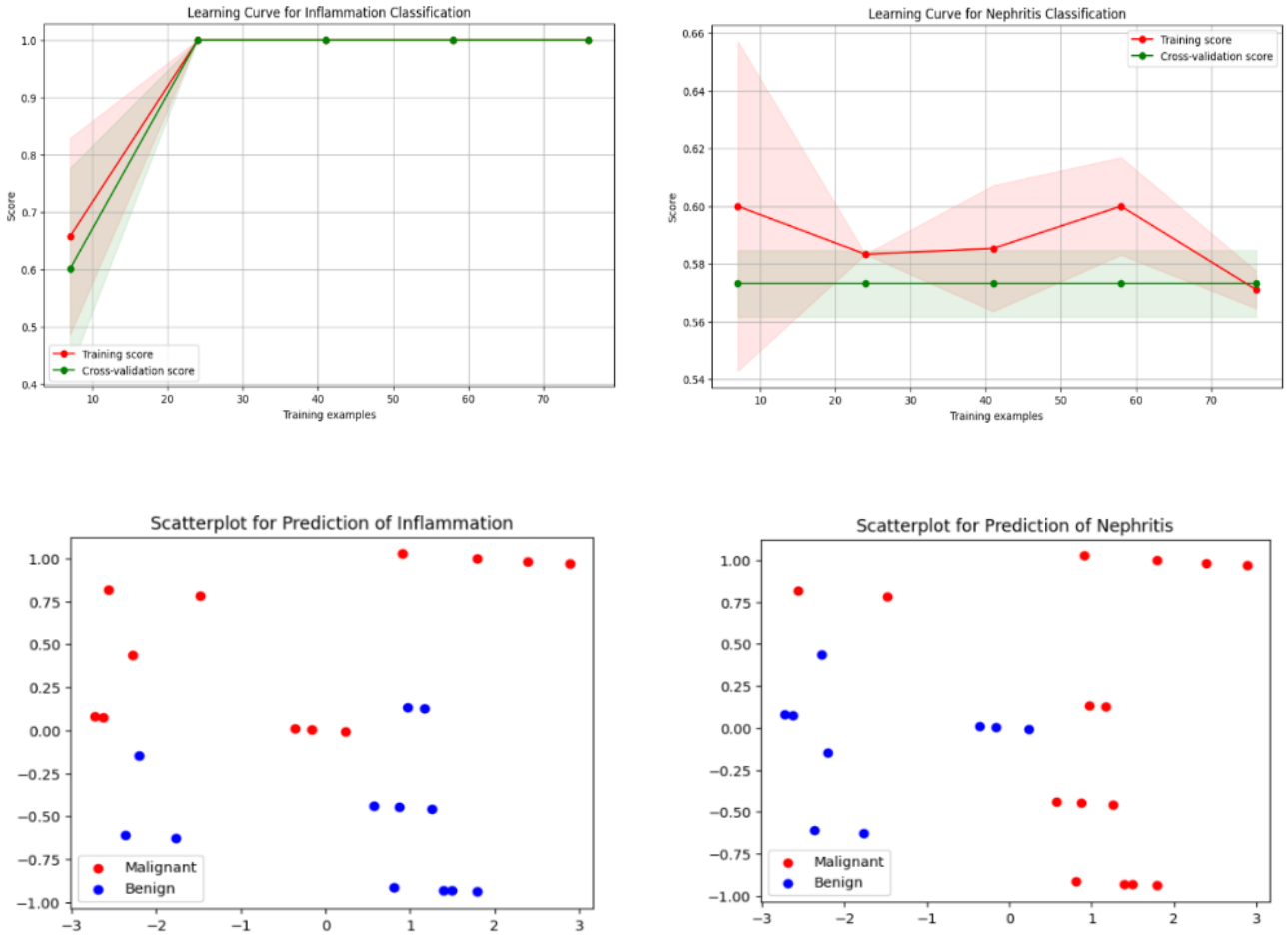


Fig 3: Results of MLPClassifier

Fig 3 represents the results of the MLP Classifier. The ANN model demonstrated exceptional accuracy (100%) and precision in predicting acute inflammation. By leveraging its capability to learn complex patterns in the data, the ANN achieved remarkable performance metrics, indicating its potential for precise classification tasks. Furthermore, learning curves provided a visualization of the models' learning behavior, showcasing their ability to generalize well to unseen data and highlighting any potential over fitting or under fitting tendencies. Scatter plots further illustrated the distribution of predicted classes, offering insights into the models' classification performance in a two-dimensional space.

Conclusion

Our study focused on predicting urinary tract infection using machine learning algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN).Utilizing the Acute inflammation dataset, we discovered valuable insights into their unique strengths and applications. Both SVM and ANN displayed exceptional accuracy, highlighting its potential for precise diagnosis of both benign and malignant cases.

By reducing the dimensionality of the dataset using PCA has enabled us to capture and retain the most important information while eliminating noise and redundancy. This dimensionality reduction enhanced model interpretability especially, when we plotted results with scatter plots image after PCA transformation, we observed a clear separation between classes, indicating improved discriminating

power of the models. Besides, learning curves showed weak convergence and reduced variance, indicating that PCA contributed to more robust reliable model performance. Overall, the results underscore the effectiveness of both SVM and ANN models in predicting acute inflammation, providing valuable insights for medical diagnosis and highlighting avenues for further research and optimization.

In summary, the aim of our work was to predict acute inflammation and nephritis using machine learning methods, specifically using Support vector machines (SVM) and multi-layer perceptron (MLP) artificial neural networks. By using the UCI Acute Inflammation dataset, we gained insight into how these algorithms work. The SVM model showed strong classification accuracy for inflammation and nephritis, which discriminated very well between positive and negative cases. Then the MLP model showed competitive performance, compared to the SVM model. Through detailed analysis and simulations, the effectiveness of these models in detecting severe urinary conditions was demonstrated. Further research could include optimizing model parameters and exploring new machine learning algorithms for improved predictive capability in clinical research projects.

Reference

[1].Kadhem, M.H. and Zeki, A.M. (2014) 'Prediction of urinary system disease diagnosis: A comparative study of three decision tree algorithms', 2014 International Conference on Computer Assisted System in Health [Preprint]. doi:10.1109/cash.2014.25.

[2].Medjahed, S. A. (2015). Urinary System Diseases diagnosis using Machine learning techniques. International Journal of Intelligent Systems and Applications, 7(5), 1–7. https://doi.org/10.5815/ijisa.2015.05.01

[3].Aleta, A., Džuho, A. and Hrvat, F. (2020) 'Detection of acute inflammation of urinary bladder and acute nephritis of renal pelvis origin using artificial neural network', 8th European Medical and Biological Engineering Conference, pp. 363–371. doi:10.1007/978-3-030-64610-3_42.

[4].Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G., & Shuaib Qureshi, M. (2023, March 14). Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease. Computational Intelligence and Neuroscience, 2023, 1–14. https://doi.org/10.1155/2023/9266889

[5].Debal, D. A., & Sitote, T. M. (2022, November 20). Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00657-5