

Efficiency measure of Machine Learning Algorithms on Liver Disease Diagnosis

Bendi Venkata Ramana

Department of IT, Aditya Institute of Technology and Management, Tekkali, A.P, India

Email: ramana.bendi@gmail.com

ABSTRACT

This The death rate in India is high due to Liver disease as a result of bad lifestyle, storage food, uncontrolled blood sugar, obesity, smoking, and consumption of alcohol and inhale of harmful gases. Earlier detection can reduce death rates and it also helps the doctors to give the proper treatment to the patients. The liver disease datasets are analyzed by using Machine learning algorithms for the accurate disease diagnosis. The datasets were collected and annotated from Visakhapatnam, Vijayawada and Tirupathi based on the major geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema respectively. Three datasets are named Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. Visakhapatnam dataset contains 12 attributes and has 499 samples. Vijayawada dataset contains 12 attributes and has 600 samples. The Tirupathi dataset contains 7 attributes and has 243 samples. The selected Classification Algorithms that are Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi-Layer Perceptron are castoff for scrutinizing their efficacy based on Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time in classifying liver patient's dataset. Classification performance is very high in the Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets, whereas Classification performance is very high in the Random Forest classification algorithm for the Vijayawada dataset. Building time is more for MLP in the Vijayawada dataset. This study motivated for the development of the Liver Diagnosis App using the Decision tree algorithm.

Keywords

Classification algorithms, liver datasets, performance

Introduction

With the increase of liver disease patients and at the same time enhanced complexity of disease diagnosis, researchers focus towards diversified machine learning algorithms for accurate identification and classification of Diseases [1]. Liver disease datasets are investigated using selected classification algorithms. The datasets considered are the Visakhapatnam dataset, Vijayawada dataset, and Tirupathi dataset based on geographical region. The selected classification algorithms considered are the naive bayes (NB) algorithm [2], decision tree (DT) algorithm [3], random forest (RF) algorithm [4], neural network (NN) algorithm [5] and support vector machines (SVM) [6]. A huge number of classification methods are used for automated liver disease diagnosis based on liver function tests (LFT). The process of liver disease classification is illustrated in Fig. 1.

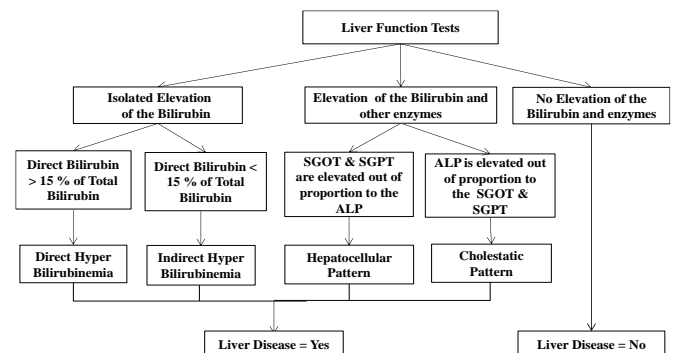


Fig. 1 Process of Liver Disease Classification

Related Work

Abbad et al. implemented KNN through distance functions in the disease diagnosis of thyroid [8]. Yao et al. proposed a densely connected deep neural network (Dense DNN) for computer aided diagnosis of Liver disease by LFT data [9]. Kuzhippallil et al. proposed an improved classification technique by integrating XGBoost and genetic algorithm. The same proposed approach is compared with other contemporary classification approaches and some other visual techniques also for the purpose of liver disease diagnosis with appropriate measuring attribute

[10]. Fathi et al. introduced SVM on ILPD and BUPA data sets for the classification of the liver and non-liver patients and presented that ILPD have maximum accuracy, sensitivity [11]. Shaheamlung et al. presented a review work by comparing some of the machine learning methods for examining and predicting liver clinical conclusions [12]. Singh et al. developed improved liver disease diagnosis forecasting system with appropriate attribute measuring rely on software paradigm to predict liver disease considering ILPD dataset [13]. Renukadevi et al. proposed an approach to resolve liver disease diagnosis by incorporating latest metaheuristics approach as grasshopper optimization algorithm by utilizing deep belief network [14]. Kumar et al. proposed Variable-NWFKNN method as an extended form of Fuzzy-NWKNN and applied over bench mark datasets considered from UCI repository [15]. Ghosh et al. evaluated Naive-Bayes, Bagging, K-Star, Logistic and REP tree rely on some performance metrics over UCLA and AP liver datasets [2]. Author Lin suggested an improved version for the purpose of liver disease analysis by combining CART and CBR methods [16]. Author Harper considered selected classification algorithms and scrutinized their efficacy and it's real world applications over various medical datasets [17]. Author Polat et al. applied Fuzzy-AIRS classification approach for the purpose of analyzing Breast Cancer and Liver problems [18-19].

Data Sets

The datasets were taken from Visakhapatnam, Vijayawada and Tirupathi based on the major geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema respectively. These datasets are examined by using machine learning methods for accurate diagnosis of liver disease and to know the impact of geographical variables such as food habits, behaviors, environment etc. on Liver Function Tests (LFT). The three datasets are named Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. Visakhapatnam dataset contains 12 attributes and has 499 samples. Vijayawada dataset contains 12 attributes and has 600 samples. Tirupathi dataset contains 7 attributes and has 243 samples. The description of datasets is given in table 1. The list and type of attributes of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset are represented in table 2, table 3 and table 4 correspondingly.

Datasets	# Attributes	# Samples	# Classes
Visakhapatnam Dataset	7	243	2
Vijayawada Dataset	12	600	2
Tirupathi Dataset	12	499	2

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
SGOT	Integer
SGPT	Integer
ALP	Integer

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
AST (SGOT)	Integer
ALT (SGPT)	Integer
ALP	Integer
IB	Real number
SP (TP)	Real number
SA (Albumin)	Real number
Globulins	Real number
A/G RATIO	Real number

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
SGOT	Integer
SGPT	Integer
ALP	Integer
IB	Real number
TP	Real number
Albumin	Real number
Globulins	Real number
A/G	Real number

Machine Learning Algorithms

Machine learning algorithms permits a system for learning with the input data as a part of making a model and it is used for predicting a given data. These methods are needed to improve the precision of models dependent on the sort and volume of the information. Machine learning algorithms are grouped into supervised, unsupervised, reinforcement and deep learning based on resemblance and learning style. These methods are used for accurate disease diagnosis and accuracy depends on the number of patient records and the learning algorithm used [20-21]. A supervised learning method learns from known input records and predicts unforeseen record. This method is categorized into one is classification and other one is regression for the development of model [22-23]. This method explores training records and iteratively forecasts class of the new record as an instructor. Classification algorithms are effectively utilized for clinical conclusions. The efficacy of such methods are examined with known records and enhancement in performance ensues with the intervention of optimization techniques [24-25]. The process of classification of disease and performance analysis is illustrated in Fig. 2.

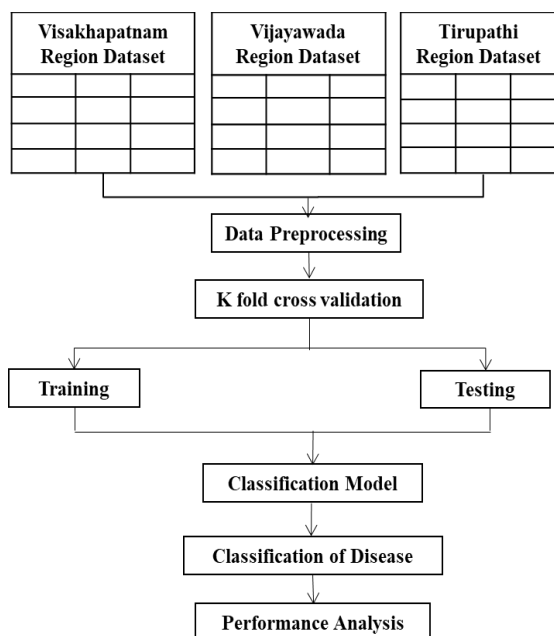


Fig. 2 Process of performance analysis

Naive-Bayes Algorithm

A naïve-Bayes (NB) classifier utilizes Bayes' probability theorem to group objects [2]. Bayes classifier utilizes likelihood hypothesis to

characterize information. Bayes' Theorem is articulated as:

$$P\left(\frac{h}{d}\right) = \frac{P\left(\frac{d}{h}\right) * P(h)}{P(d)}$$

Where P(h/d) is the likelihood of theory h given the information d. This is known as the back likelihood. P(d/h) is the likelihood of information d given that the speculation h was valid. P(h) is the likelihood of theory h being valid. This is known as the earlier likelihood of h. P(d) is the likelihood of the information.

Decision Tree Algorithm

The structure in Decision Tree is as a normal tree comprises root, branches and leaves. Each node in this tree illustrates an attribute, each link illustrates a decision and each leaf illustrates a conclusion. Decision Tree is analogous to the decision-making done by the human. It can resolve both discrete and continuous data. As a technique, it permits you to move towards the issue in an organized and methodical manner to come to a coherent end result [3].

Random Forest Algorithm

Random forest (RF) falls to the category of machine learning technique which is utilized to resolve classification and regression issues as this belongs to supervised approach. The process initiates by choosing of samples randomly from selected dataset and construction of decision tree starts by utilizing the algorithm for each sample, expecting proper predicted outcome. Along with the predicted outcome voting will be accomplished. Finally best voted outcome is considered as the optimum [4]. To achieve the best split using RF algorithm appropriate features are considered randomly. At each split position attributes to be examined, are signified as one input to this algorithm.

Neural Network Algorithm

Feed forward neural networks (FFNN) were among the first and simple approach for resolving non-linear complications. The meaning of feed forward means it will go in one direction only. Various types of FFNNs are available and popularly used by research community starting from simple MLPs to other higher order neural networks (HONN) including deep networks [5]. Generally all networks comprises of minimum one input, one

hidden and one output layer. These layers are interconnected with the help of neurons associated with weights. For the mapping of non-linear functions activation function is utilized to deal with. At present HONNs are mostly utilized to deal with complicated real world problems [26-29].

Support Vector Machines

Support vector machines (SVM) are one of the well-known machine learning technique utilized to resolve prediction, classification and regression complications. The fundamental goal of SVM is to locate the ideal hyper plane which straightly isolates the information focuses in two segments by optimizing the boundary. SVM is most suitable to apply on natural language processing tasks due to lack of mostly complex stuff [6-7].

Performance Evaluation

Due to the varying nature of the problems, one classifier is not suitable to solve all kinds of complications. Based on this, different classifiers are evolved day by day and its efficiency has been scrutinized by various means. Here the performance of different machine learning classifiers are examined by virtue of percentage of correct and incorrect classified samples in terms of training and testing. For the same purpose confusion matrix has been considered for examining the efficacy of the classifier, depending on suitable datasets. In this study binary class problem has been considered, and the confusion matrix is made up of four conclusions. The performance metrics is used to test the effectiveness of classifiers are accuracy, precision, sensitivity, specificity, false positive rate, F-measure, ROC curve, mean absolute error, root mean squared error, root relative squared error, kappa statistics and building time. The confusion matrix and performance metrics for castoff classifiers are presented in Fig. 3 and Fig. 4 correspondingly.

Predicted Values → Actual Values ↓	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Fig. 3 Confusion matrix

Results and Discussion

In this experimentation, considered classifiers are utilized for the assessment of diagnosis of liver disease. The classifiers considered in this study are

NB algorithm, DT algorithm, RF algorithm, SVM and MLP. The three liver datasets from various regions of Andhra Pradesh were considered for the evaluation of classification algorithms based on the various performance evaluators. The performance evaluators are Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time. For the validation purpose considered technique is 10-fold cross validation. This the entire dataset is subdivided into 10 equal portions, from which nine portions are considered for the training purpose and rest one portion is utilized for testing purpose. Reiterating such ten portions signifies that entire portions are used for the sake of testing and training to reduce sample bias. Efficiency measures and error measures are evaluated for the NB, DT, RF, SVM and MLP on Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset. These measures are depicted in table 5. Performance comparison of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset for the NB, DT, RF, SVM and MLP are depicted in Fig. 5, Fig. 7 and Fig. 9 respectively. Error comparison of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset for the NB, DT, RF, SVMs and MLP are reported in Fig. 6, Fig. 8 and Fig. 10 respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2 * Precision + Recall}{Precision + Recall}$$

$$MAE = \frac{1}{n} \sum |y_i - y^j|$$

$$RMSE = \sqrt{(f - o) * 2}$$

$$RRSE = \frac{RMSE}{RMPSE}$$

$$K = \frac{Observed\ agreement - chance\ agreement}{1 - chance\ agreement}$$

TP = True Positives
 TN = True Negatives
 FP = False Positives
 FN = False Negatives
 FPR = False Positive Rate
 MAE = Mean Absolute Error
 RMSE = Root Mean Squared Error
 RMPSE = Root Mean Prior Squared Error
 RRSE = Root Relative Squared Error
 f = forecasts (expected values or unknown results),
 o = observed values (known results)
 BT = Building Time (Time required to build classifier)
 K = Kappa Statics

Fig. 4 Performance metrics for classifiers

Accuracy, Precision, Sensitivity and Specificity are very high in Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets and subsequently error measures are very less for the same datasets. Accuracy, Precision, Sensitivity and Specificity are very high in Random Forest classification algorithm for Vijayawada dataset.

Building time is more for MLP than other classifiers in all the three datasets. Building time for MLP in Vijayawada dataset is more than Visakhapatnam and Tirupathi dataset. This may be due to more no of records in Vijayawada dataset than other datasets.

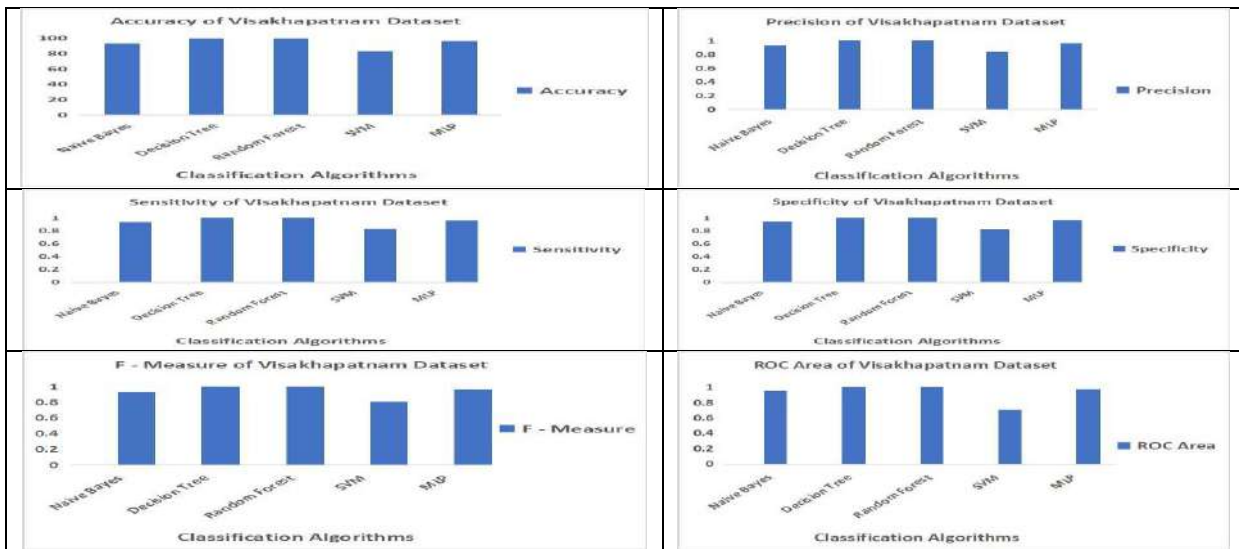


Fig. 5 Performance comparison of Visakhapatnam dataset

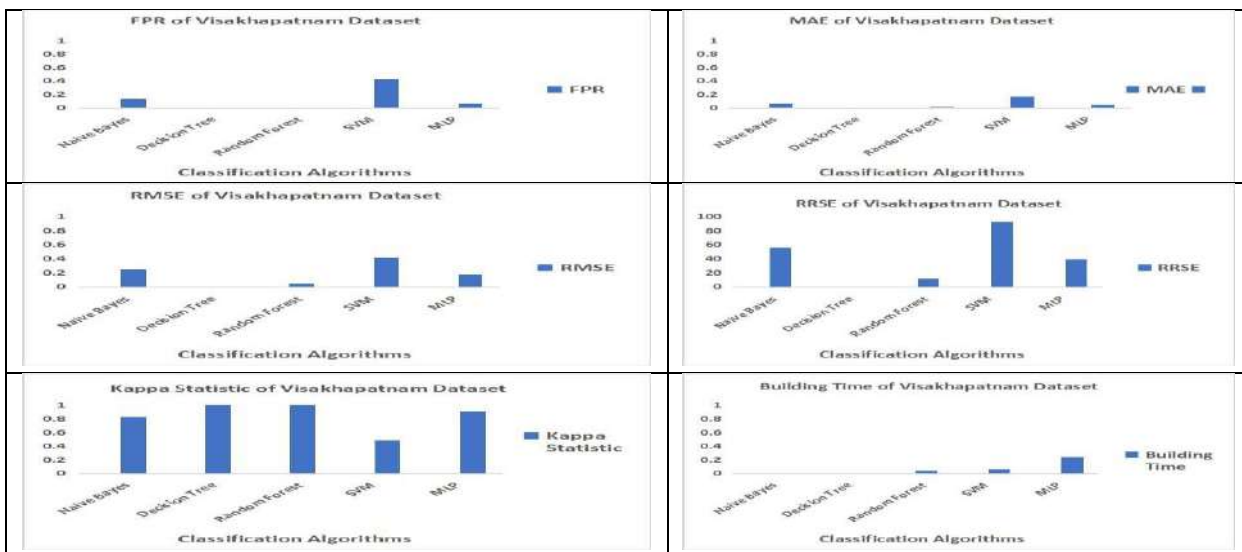
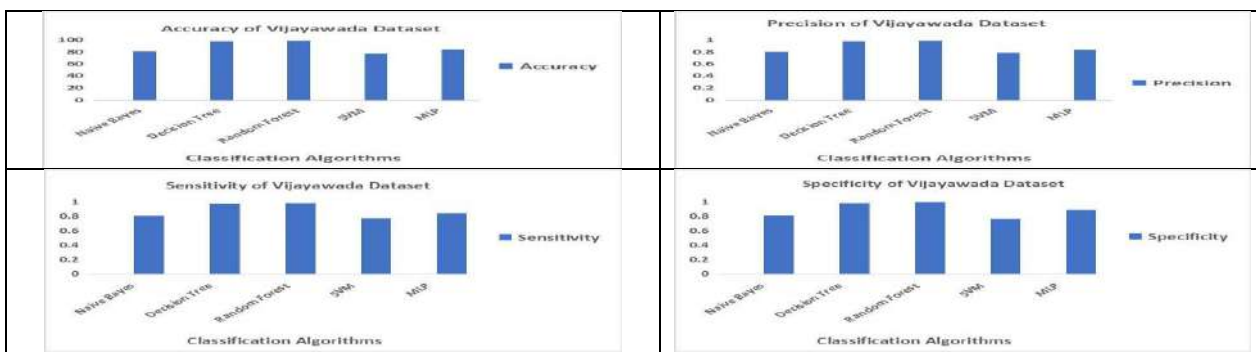


Fig. 6 Error comparison of Visakhapatnam dataset



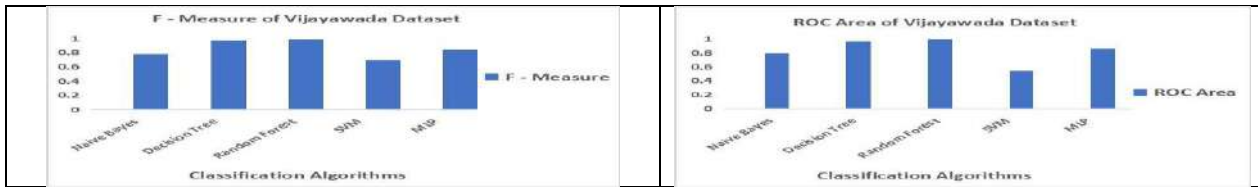


Fig. 7 Performance comparison of Vijayawada dataset

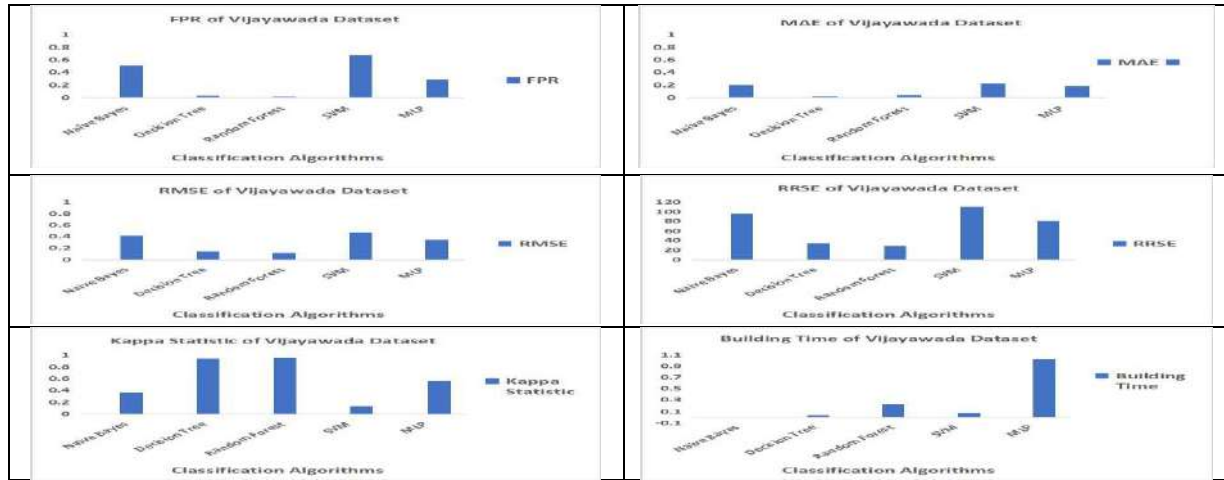


Fig. 8 Error comparison of Vijayawada dataset

<i>Datasets \ Algorithm</i>		<i>Naive Bayes</i>	<i>Decision Tree</i>	<i>Random Forest</i>	<i>SVM</i>	<i>MLP</i>
Visakha patnam Dataset	Accuracy	93.4156	100	100	83.1276	96.707
	Precision	0.934	1.000	1.000	0.838	0.967
	Sensitivity	0.934	1.000	1.000	0.831	0.967
	Specificity	0.9402	1.000	1.000	0.8246	0.9722
	F-Measure	0.933	1.000	1.000	0.810	0.967
	ROC-Area	0.951	1.000	1.000	0.704	0.978
	FPR	0.131	0.000	0.000	0.423	0.061
	MAE	0.0641	0	0.0127	0.1687	0.0526
	RMSE	0.2473	0	0.0519	0.4108	0.1747
	RRSE	55.857	0	11.7123	92.7776	39.454
	Kappa Statistic	0.8269	1	1	0.4867	0.9152
Building Time (Sec)	0	0	0.04	0.06	0.24	
Vijayawada Dataset	Accuracy	80.8333	97.6667	98.5	77.1667	84.333
	Precision	0.800	0.977	0.986	0.786	0.839
	Sensitivity	0.808	0.977	0.985	0.772	0.843
	Specificity	0.8146	0.9866	0.9977	0.7697	0.8896
	F-Measure	0.779	0.977	0.985	0.697	0.841
	ROC-Area	0.790	0.964	0.993	0.547	0.865
	FPR	0.508	0.035	0.009	0.678	0.294
	MAE	0.198	0.0246	0.0438	0.2283	0.1798
	RMSE	0.4169	0.1467	0.1226	0.4778	0.3506
	RRSE	96.5015	33.9636	28.3694	110.597	81.142
Kappa Statistic	0.3689	0.9378	0.9604	0.1332	0.5667	

	Building Time (Sec)	0	0.03	0.23	0.08	1.02
Tirupathi Dataset	Accuracy	98.5972	100	99.7996	99.5992	99.599
	Precision	0.986	1.000	0.998	0.996	0.996
	Sensitivity	0.986	1.000	0.998	0.996	0.996
	Specificity	0.9952	1.000	0.9953	0.9907	0.9953
	F-Measure	0.986	1.000	0.998	0.996	0.996
	ROC-Area	1.000	1.000	0.999	0.996	0.995
	FPR	0.018	0.000	0.002	0.003	0.004
	MAE	0.0365	0	0.0146	0.004	0.006
	RMSE	0.1118	0	0.0517	0.0633	0.0641
	RRSE	22.5796	0	10.4514	12.7914	12.943
	Kappa Statistic	0.9713	1	0.9959	0.9918	0.9918
	Building Time (Sec)	0.01	0.03	0.24	0.15	0.9

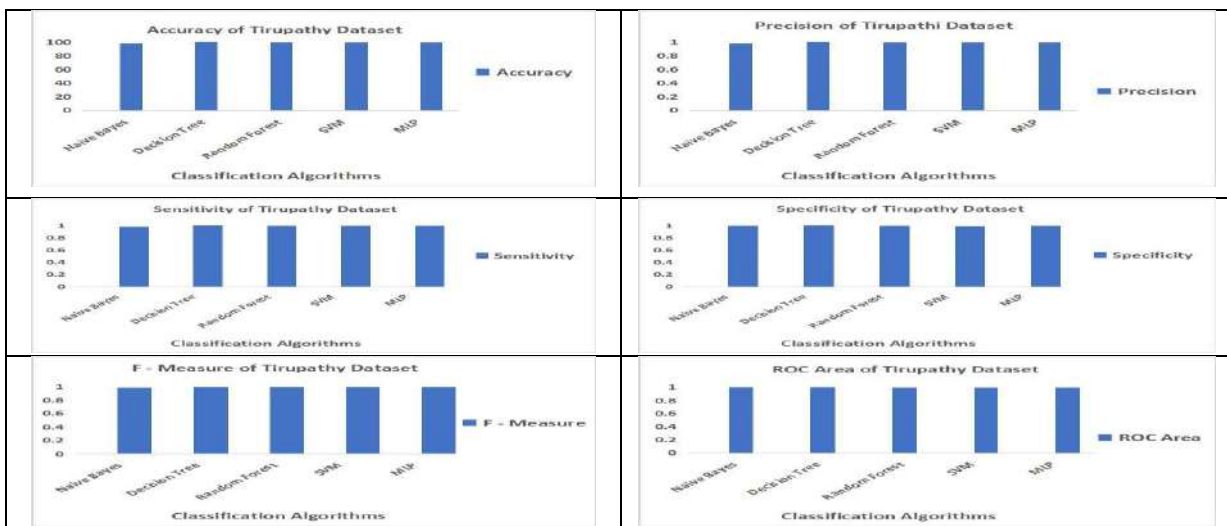


Fig. 9 Performance comparison of Tirupathi dataset

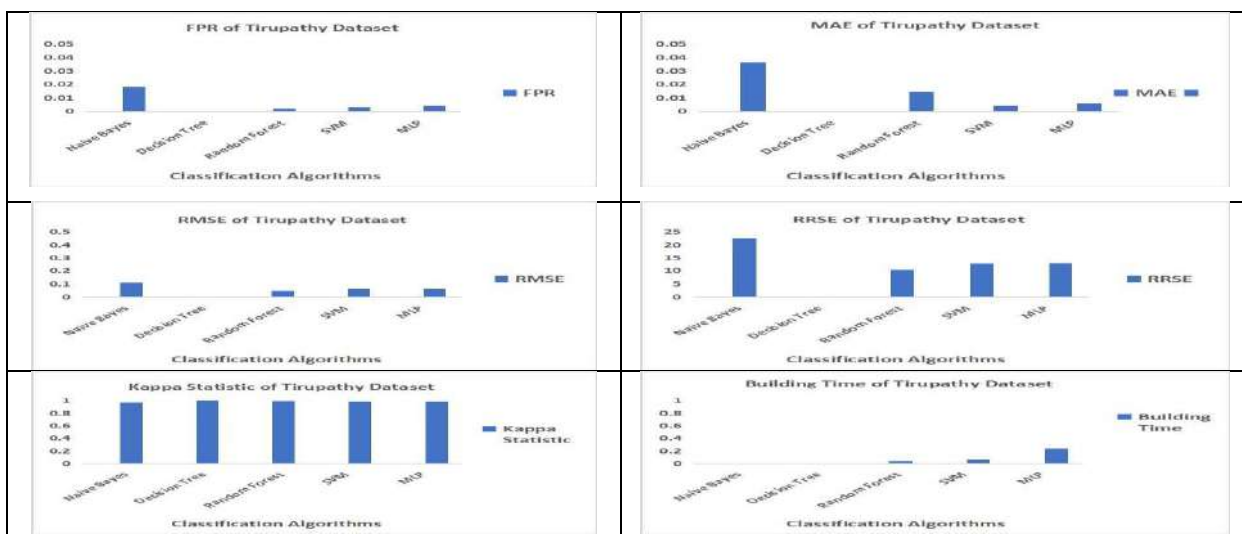


Fig. 10 Error comparison of Tirupathi dataset

Conclusions

In this experimentation, Naive Bayes, Decision Tree, Random Forest, SVM and Multi-Layer Perceptron classification techniques has been considered for assessing performance efficacy and represented by considering measures such as Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time in classifying liver patients dataset. Classification performance is very high in Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets, whereas Classification performance is very high in Random Forest classification algorithm for Vijayawada dataset. Building time is more for MLP in Vijayawada dataset.

Future Scope

The performance of classification algorithms may be improved by selecting important features in classification of liver disease diagnosis. It can also be enhanced by ensembling the classifiers. The performance of classification algorithm is further improved by Optimization algorithms.

Acknowledgments

I am thankful to UGC, New Delhi for granting financial support under UGC Minor Research Project scheme with reference no. MRP-6934/16 (SERO/UGC). I am also thankful to Prof. V. V. Nageswara Rao, Director of Aditya Institute of Technology and Management (A) for providing necessary facilities.

References

- [1] Panda, N. and Majhi, S.K., 2020. Improved salp swarm algorithm with space transformation search for training neural network. *Arabian Journal for Science and Engineering*, 45(4), pp.2743-2761.
- [2] Ghosh, S.R. and Waheed, S., 2017. Analysis of classification algorithms for liver disease diagnosis. *Journal of Science, Technology and Environment Informatics*, 5(01), pp.361-370.
- [3] Macchiavello, G., Moser, G., Boni, G. and Serpico, S.B., 2009, July. Automatic unsupervised classification of snow-covered areas by decision-tree classification and minimum-error thresholding. In *2009 IEEE International Geoscience and Remote Sensing Symposium (Vol. 2, pp. II-1000)*. IEEE.
- [4] Izquierdo-Verdiguier, E. and Zurita-Milla, R., 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 88, p.102051.
- [5] Panda, N. and Majhi, S.K., 2020. How effective is the salp swarm algorithm in data classification. In *Computational Intelligence in Pattern Recognition (pp. 579-588)*. Springer, Singapore.
- [6] Ren, L., Chang, H. and Yi, Y., 2009, July. An Improved Binary Tree SVM Classification Algorithm Based on Bayesian. In *2009 Asia-Pacific Conference on Information Processing (Vol. 1, pp. 178-181)*. IEEE.
- [7] Mathur, A. and Foody, G.M., 2008. Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2), pp.241-245.
- [8] Abbad Ur Rehman, H., Lin, C.Y. and Mushtaq, Z., 2021. Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. *Journal of the Chinese Institute of Engineers*, 44(1), pp.77-87.
- [9] Yao, Z., Li, J., Guan, Z., Ye, Y. and Chen, Y., 2020. Liver disease screening based on densely connected deep neural networks. *Neural Networks*, 123, pp.299-304.
- [10] Kuzhippallil, M.A., Joseph, C. and Kannan, A., 2020, March. Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 778-782)*. IEEE.
- [11] Fathi, M., Nemati, M., Mohammadi, S.M. and Abbasi-Kesbi, R., 2020. A machine learning approach based on SVM for classification of liver diseases. *Biomedical Engineering: Applications, Basis and Communications*, 32(03), p.2050018.
- [12] Shaheamlung, G., Kaur, H. and Kaur, M., 2020, June. A Survey on machine learning techniques for the diagnosis of liver disease. In *2020 International Conference on Intelligent Engineering and Management (ICIEM) (pp. 337-341)*. IEEE.
- [13] Singh, J., Bagga, S. and Kaur, R., 2020. Software-based Prediction of Liver Disease

- with Feature Selection and Classification Techniques. *Procedia Computer Science*, 167, pp.1970-1980.
- [14] Renukadevi, T. and Karunakaran, S., 2020. Optimizing deep belief network parameters using grasshopper algorithm for liver disease classification. *International Journal of Imaging Systems and Technology*, 30(1), pp.168-184.
- [15] Kumar, P. and Thakur, R.S., 2020. Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach. *Multimedia Tools and Applications*, pp.1-21.
- [16] Lin, R.H., 2009. An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, 47(1), pp.53-62.
- [17] Harper, P.R., 2005. A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71(3), pp.315-331.
- [18] Kahramanli, H. and Allahverdi, N., 2009. Mining Classification Rules for Liver Disorders. *International journal of mathematics and computers in simulation*, 3(1), pp.9-19.
- [19] Polat, K., Şahan, S., Kodaz, H. and Güneş, S., 2007. Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. *Expert Systems with Applications*, 32(1), pp.172-183.
- [20] Veena, K.M., Shenoy, K.M. and Shenoy, K.A., 2018, April. Performance comparison of machine learning classification algorithms. In *International Conference on Advances in Computing and Data Sciences* (pp. 489-497). Springer, Singapore.
- [21] Biagetti, G., Crippa, P., Falaschetti, L., Tanoni, G. and Turchetti, C., 2018. A comparative study of machine learning algorithms for physiological signal classification. *Procedia computer science*, 126, pp.1977-1984.
- [22] Abdullah, N.E., Hashim, H., Osman, F.N. and Adam, F.M., 2010, April. Comparison between various supervised ANN algorithm using RGB indices for plaque lesion classification. In *2010 International Conference on Electronic Devices, Systems and Applications* (pp. 236-241). IEEE.
- [23] Ramana, B.V. and Boddu, R.S.K., 2019, January. Performance comparison of classification algorithms on medical datasets. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0140-0145). IEEE.
- [24] Firouzabadi, P.Z., 2001, July. Performance evaluation of supervised classification of remotely sensed data for crop acreage estimation. In *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217)* (Vol. 6, pp. 2718-2720). IEEE.
- [25] Cufoglu, A., Lohi, M. and Madani, K., 2009, March. A comparative study of selected classifiers with classification accuracy in user profiling. In *2009 WRI World Congress on Computer Science and Information Engineering* (Vol. 3, pp. 708-712). IEEE.
- [26] Panda, N. and Majhi, S.K., 2020. Effectiveness of Swarm-Based Metaheuristic Algorithm in Data Classification Using Pi-Sigma Higher Order Neural Network. In *Progress in Advanced Computing and Intelligent Engineering* (pp. 77-88). Springer, Singapore.
- [27] Panda, N. and Majhi, S.K., Oppositional salp swarm algorithm with mutation operator for global optimization and application in training higher order neural networks. *Multimedia Tools and Applications*, pp.1-25.
- [28] Panda, N., Majhi, S.K., Singh, S. and Khanna, A., 2020. Oppositional spotted hyena optimizer with mutation operator for global optimization and application in training wavelet neural network. *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-14.
- [29] Panda, N. and Majhi, S.K., 2020. Improved spotted hyena optimizer with space transformational search for training pi-sigma higher order neural network. *Computational Intelligence*, 36(1), pp.320-350.

Analysis of Geographical effect of various regions on Liver disease

Bendi Venkata Ramana

Department of IT, Aditya Institute of Technology and Management, Tekkali, A.P, India

Email: ramana.bendi@gmail.com

ABSTRACT

Statistical Analysis plays a significant role in population comparison to investigate the geographical effect on liver diseases. In this study the common attributes ALP, DB, SGOT, SGPT and TB were considered from the three datasets for the population comparison. Three data sets were assessed using analysis of variance and multivariate analysis of variance and significance level observed for the statistical analysis is ≤ 0.05 for the corresponding confidence level is 95%. The Significant values in the ANOVA and MANOVA analysis indicates there is more significant difference among three liver datasets that means there is a geographical effect on liver diseases.

Keywords

Statistical Analysis, liver datasets, population comparison

Introduction

Statistical Analysis plays a significant role in population comparison to investigate the geographical effect on liver diseases. In this study the common attributes were considered from the three datasets for the population comparison. Three data sets were assessed using analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). ANOVA analysis applied on each single dependent variable that is ALP, DB, SGOT, SGPT and TB to test the significant difference among three datasets. MANOVA statistical analysis [1-5] applied on all combinations of common attributes to envisage the significance among three datasets.

Related Work

Wong et al. evaluated hepatic fibrosis among Nonalcoholic fatty liver disease patients, concentrating on dietary patterns [6]. Xia et al. investigated rate of Nonalcoholic fatty liver disease regional wise among liver fat content and glucose metabolism [7]. Cannon et al. justified exceptions other than hepatocellular carcinoma for final stage liver disease [8]. Jothimani et al. reported that the mortality rate is more for liver patients with COVID-19 [9]. Author Liao et al. explored the correlation among medical-level factors, county-level factors and rate of antibiotic use [10]. Khorraminezhad et al. explored the approaches for the analysis of multi-OMICs data in nutrition studies [11]. Bendi, V.R. et al. investigated performance of selected classification algorithms

on medical datasets taken from the UCI repository [12]. Percival et al. analyzed statistically using multicomponent nuclear magnetic resonance [13]. Yousaf et al. forecasted COVID-19 confirmed cases and death cases using the Auto-Regressive Integrated Moving Average Model [14]. Ustaoglu et al. accomplished statistical analysis using pearson correlation index, principal component and clustering analysis to determine water quality of the stream [15]. Petersen et al. performed statistical analysis on many patients increasing hypoxic respiratory failure due to COVID-19 [16]. Gyimah et al. investigated statistically using principal component and cluster analyses for the assessment of Densu River water quality [17]. Kattan et al. explored papers that are submitted to CHEST will be reviewed using the statistical analysis and give response to peer review [18]. Ren et al. accomplished statistical analysis related to tunnel fire accidents based on causes, characteristics, and consequences [19]. Author Livadiotis implemented a statistical analysis on progress rate of COVID-19 cases with the effect of temperature [20]. Verma et al. analyzed statistically using KNN and K-means on CIDDS-001 dataset for assessing Network Intrusion Detection Systems [21]. Zajkowska et al. performed multivariate statistical analysis for diagnosing breast cancer at an initial and final stage [22]. Mustaqeem et al. developed a model that predicts cardiac disease patients by assessing clinical features of patients [23]. Zhao et al. examined transmission chain of secondary cases COVID-19 [24]. Lara-Cabrera et al. considered the efficacy of metrics to assess the risk of

radicalization [25]. Barbulescu et al. considered the perceptions dependent on the intensity of dust storm every month, season, year and location [26]. Sidorov et al. deliberated the practical approach for processing the research data by using research techniques and statistics [27]. Nordhausen et al. anticipated Independent component analysis for the refinement of principal component analysis by using covariance matrix [28].

Data Sets

Datasets were taken from Visakhapatnam, Vijayawada and Tirupathi based on the major geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema respectively. These datasets [21] are analyzed by using statistical techniques to know the impact of geographical variables such as food habits, behaviors, environment etc on Liver Function Tests (LFT) [22-23]. The three datasets are named Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. Visakhapatnam dataset contains 12 attributes and has 499 samples. Vijayawada dataset contains 12 attributes and has 600 samples. Tirupathi dataset contains 7 attributes and has 243 samples. The description of datasets is given table1.

Datasets	# Attributes	# Samples	# Classes
Visakhapatnam Dataset	7	243	2
Vijayawada Dataset	12	600	2
Tirupathi Dataset	12	499	2

Statistical Analysis

Statistical Analysis involves the investigation of techniques for gathering, summing up, and deciphering information. Measurements formalize the way toward deciding. The uses of measurements in sciences, financial aspects, software engineering, money, brain research, humanism, criminology, and numerous different fields. It will inspect various approaches to research the connections between different attributes of information. In this investigation standard statistical techniques [25] ANOVA and MANOVA are applied to assess the importance between two populaces for better grouping. ANOVA is utilized to test the considerable difference between one dependent variable and one independent variable. MANOVA is utilized to test the considerable difference between more dependent variable and individual independent variable. The truth liver datasets were taken from Visakhapatnam, Vijayawada and Tirupathi based on the main geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema region respectively. The attributes in these datasets are represented in the table 2. The common attributes from the three data sets AGE, GENDER, TB, DB, SGOT, SGPT and ALP are considered for the purpose of population comparison to analyze the geographical effect. In this Group 1 indicates Visakhapatnam dataset, Group 2 indicates Vijayawada dataset and Group 3 indicates Tirupathi dataset.

Table 2. List of attributes and types of regional datasets

<i>Visakhapatnam Region</i>		<i>Vijayawada Region</i>		<i>Tirupathi Region</i>	
<i>Attribute</i>	<i>Type</i>	<i>Attribute</i>	<i>Type</i>	<i>Attribute</i>	<i>Type</i>
AGE	Real number	AGE	Real number	AGE	Real number
GENDER	Categorical	GENDER	Categorical	GENDER	Categorical
TB	Real number	TB	Real number	TB	Real number
DB	Real number	DB	Real number	DB	Real number
SGOT	Integer	AST (SGOT)	Integer	SGOT	Integer
SGPT	Integer	ALT (SGPT)	Integer	SGPT	Integer
ALP	Integer	ALP	Integer	ALP	Integer
		IB	Real number	IB	Real number
		SP (TP)	Real number	TP	Real number
		SA (Albumin)	Real number	Albumin	Real number
		Globulins	Real number	Globulins	Real number
		A/G RATIO	Real number	A/G RATIO	Real number

Analysis of variance

ANOVA is used to analyze statistically [14-16] between one dependent variable and two or more categories. In this between groups directs variability based on place of data, within groups directs variability based on random error and total directs total variability. F-statistic is a ratio of variation between group and within group. ANOVA and MANOVA statistical techniques are considered for the population comparison between datasets to know the geographical effect on liver datasets. Statistical analysis metrics depicted in Fig. 1.

$$SS_{between} = n \sum (\bar{x} - \bar{x}_{grand})^2$$

$$SS_{within} = \sum (x - \bar{x}_{group})^2$$

$$F - \text{statistic} = \frac{\text{Between Group Variation}}{\text{Within Group Variation}}$$

$$X_{grand\ mean} = \frac{\sum \bar{x}}{N}$$

$$X_{group\ mean} = \frac{\sum fm}{\sum f}$$

$SS_{between}$ = Sum of the squares between groups
 SS_{within} = Sum of the squares within Groups
 N = Total number of sets
 $\sum \bar{x}$ = Sum of mean of all sets
 $df_{total} = N - 1$
 $df_{between} = K - 1$
 $df_{within} = N - K$
 N = Number of all scores
 K = Number of groups
 Degrees of freedom = number of values estimated by the statistical analysis
 f = frequency
 m = midpoint
 X = Sample value
 \bar{x} = Mean of the sample

Fig. 1 Statistical Analysis Metrics

Multivariate Analysis of variance

MANOVA is utilized to assess the hypothesis that at least one independent factors (IVs), affect a group of at least two dependent factors (DVs). The objective of our investigation is to search for an impact of at least one IVs on a few DVs simultaneously. Four distinctive multivariate tests [17] were considered to distinguish the huge impact of the IVs on the entirety of the DVs, as a gathering. Descriptive statistics are categorized into two measures. One is central tendency that comprise mean, median, and mode. Another one is dispersion of data that comprises standard deviation, variance, the minimum and maximum.

Results and Discussion

The common features ALP, DB, SGOT, SGPT and TB were considered as DVs and group was considered as factor variable. The consequences of ANOVA were addressed in three lines. The outcome of ANOVA analysis on ALP, DB, SGOT, SGPT and TB was addressed in table 3, table 4, table 5, table 6 and table 7 correspondingly which demonstrates whether there is a measurably difference among means of groups.

Table. 3: ANOVA on ALP between three datasets

<i>ANOVA - ALP</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	382.012	226	1.690	6.057	.000
Within Groups	311.154	1115	.279		
Total	693.165	1341			

<i>ANOVA - DB</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	279.482	46	6.076	19.019	.000
Within Groups	413.684	1295	.319		
Total	693.165	1341			

<i>ANOVA - SGOT</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	257.506	127	2.028	5.650	.000
Within Groups	435.659	1214	.359		
Total	693.165	1341			

<i>ANOVA - SGPT</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	179.806	129	1.394	3.288	.000
Within Groups	513.323	1211	.424		
Total	693.129	1340			

<i>ANOVA - SGPT</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Between Groups	141.279	61	2.316	5.372	.000
Within Groups	551.886	1280	.431		
Total	693.165	1341			

Significance value indicates the possibility of getting a change in mean among the groups as high as seen by chance. More considerable change among the groups due to the lower p-value. In this investigation the sig. value is less than 0.05 means there will be more considerable variance among groups. This specifies the groups have considerable variances on ALP, DB, SGOT, SGPT and TB. ANOVA explores statistically considerable change

totally among the groups and there is need of MANOVA analysis for Multiple Comparisons, to explore which groups differed from each other. Descriptive Statistics of Visakhapatnam dataset, descriptive Statistics of Vijayawada dataset and descriptive Statistics of Tirupathi dataset are represented in table 8, table 9 and table 10 correspondingly.

<i>Attribute</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Mean</i>	<i>Standard Deviation</i>
AGE	5	68	29.207	12.226
TB	0.3	18.2	1.474	2.172
DB	0.1	13	0.604	1.514
SGOT	14	175	40.626	26.324
SGPT	10	267	36.247	32.343
ALP	28	605	95.753	87.995

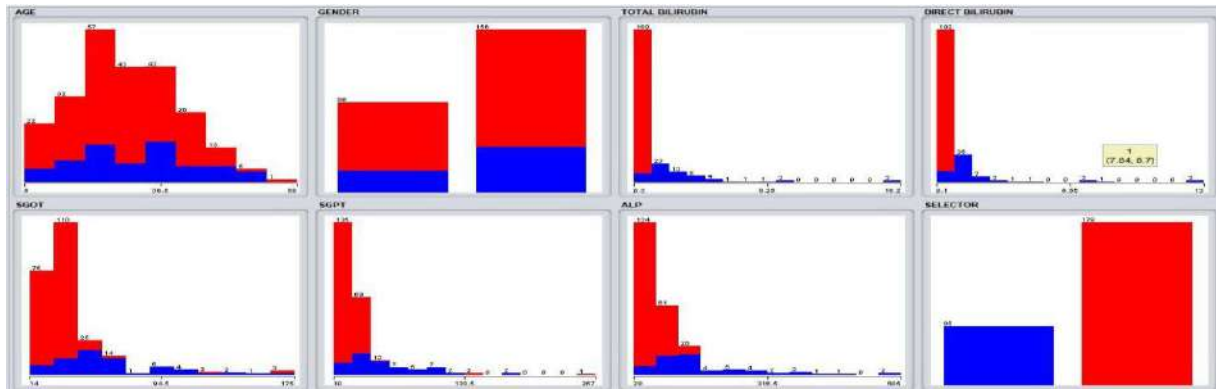


Fig. 1. Descriptive Statistics of Visakhapatnam Dataset

Attribute	Minimum	Maximum	Mean	Standard Deviation
AGE	0	90	45.748	17.782
TB	0.1	31.6	0.925	1.886
DB	0.1	22.8	0.447	1.433
SGOT	2	4980	58.523	244.4
SGPT	2	3642	57.552	220.689
ALP	0.7	769	114.74	90.394
IB	0.1	15	0.511	0.979
TP	2.3	502	7.842	20.523
Albumin	0.8	4.9	3.147	.0754
Globulins	12	24	3.735	1.37
A/G RATIO	0.1	2.5	0.903	0.321

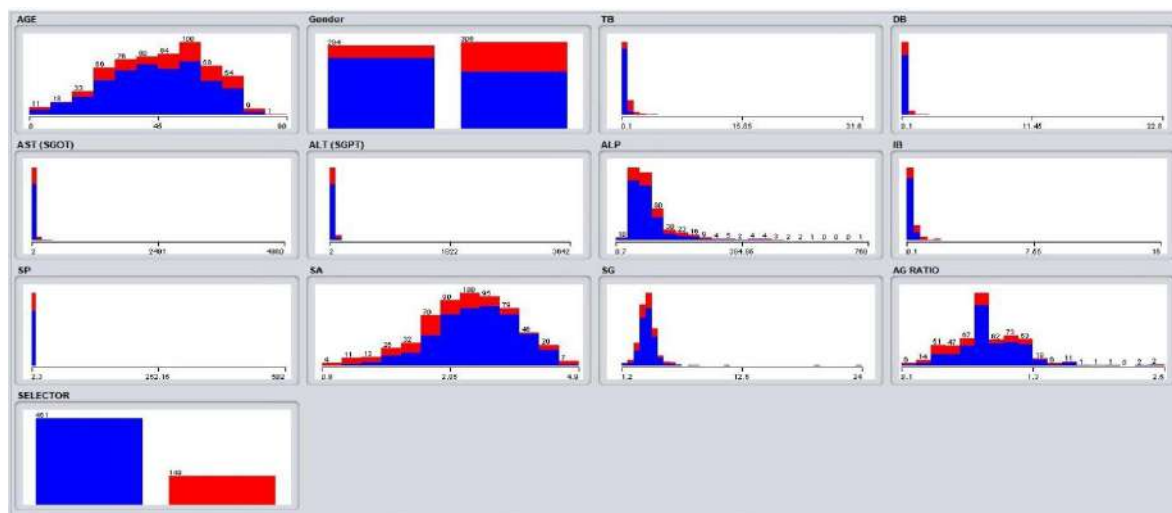


Fig. 2. Descriptive Statistics of Vijayawada Dataset

Attribute	Minimum	Maximum	Mean	Standard Deviation
AGE	19	85	47.259	13.905
TB	0.2	1	0.726	0.138
DB	0.1	0.5	0.257	0.054
SGOT	10	45	23.733	7.97
SGPT	11	45	20.657	6.27
ALP	18	188	90.717	19.38

IB	0.1	0.7	0.468	0.128
TP	5.8	72	7.331	2.927
Albumin	2.9	4.8	4.17	0.293
Globulins	2.1	25	3.09	1.053
A/G RATIO	1	2.2	1.365	0.239

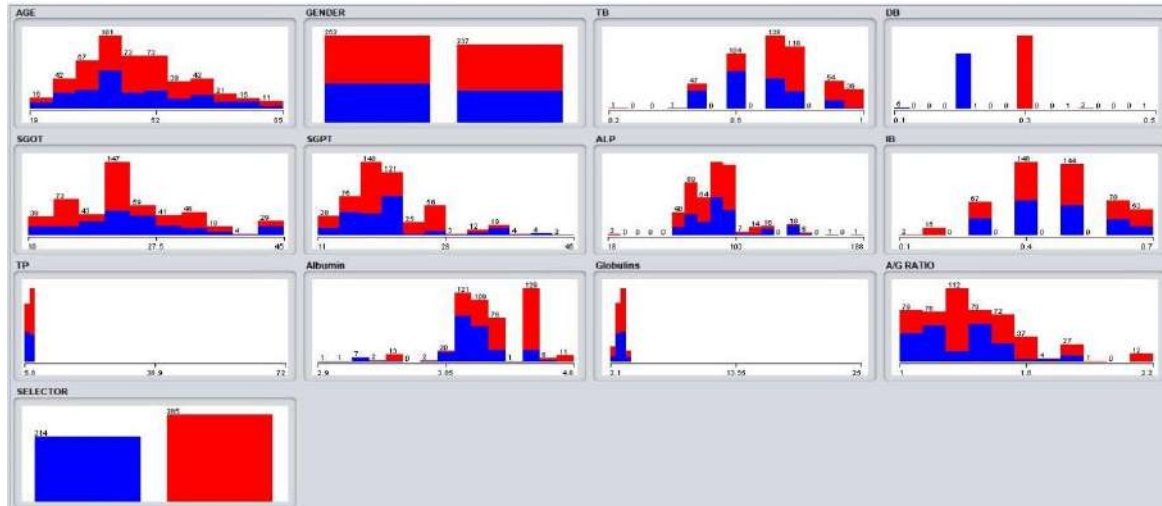


Fig. 3. Descriptive Statistics of Tirupathi Dataset

Multivariate analysis offers more substitute statistical tests and whereas one statistical test available in ANOVA. It has Wilks’ Lambda, Lawley’s trace, Pillai’s trace and Roy’s largest root statistical tests. They have generated same results with the hypothesis degrees of freedom is 1 and whereas for the degrees of freedom is greater than 1, Wilks’ Lambda, Lawley’s trace, and Roy’s largest root are more powerful than Pillai’s trace. MANOVA studies the degree of variance within

the IVs and concludes the degree of variance between the IVs. IVs had a significant effect on the DV in within subjects variance and smaller than the between subjects variance. In this study the common attributes from the three datasets TB, DB, SGOT, SGPT and ALP are considered for the multivariate analysis. Multivariate assessments for the blend of attributes at the respective significant values at diverse levels are denoted in Table 11.

Table 11. Multivariate Tests Significance (at p < 0.05 level)

<i>Level</i>	<i>variables</i>	<i>Pillai's Trace Value</i>	<i>F</i>	<i>Hypothesis df</i>	<i>Error df</i>	<i>Sig.</i>	<i>Partial Eta Squared</i>
2-way Interactions	DB*ALP	.629	1133.53	2.000	1338.000	.000	.629
	DB*SGOT	.135	104.031	2.000	1338.000	.000	.135
	DB*SGPT	.137	106.212	2.000	1337.000	.000	.137
	SGOT*ALP	.628	1127.60	2.000	1338.000	.000	.628
	SGPT*ALP	.628	1126.30	2.000	1337.000	.000	.628
	SGOT*SGPT	.065	46.306	2.000	1337.000	.000	.065
	TB*ALP	.630	1138.87	2.000	1338.000	.000	.630
	TB*DB	.399	443.812	2.000	1338.000	.000	.399
	TB*SGOT	.286	268.204	2.000	1338.000	.000	.286
	TB*SGPT	.288	270.857	2.000	1337.000	.000	.288
3-way Interactions	DB*SGOT*ALP	.631	761.415	3.000	1337.000	.000	.631
	DB*SGOT*SGPT	.142	73.857	3.000	1336.000	.000	.142

ons	SGOT*SGPT*ALP	.628	751.625	3.000	1336.000	.000	.628
	TB*DB*SGOT	.407	306.048	3.000	1337.000	.000	.407
	TB*DB*SGPT	.410	308.974	3.000	1336.000	.000	.410
4-way Interactions	DB*SGOT*SGPT*ALP	.631	571.128	4.000	1335.000	.000	.631
	TB*DB*SGOT*ALP	.682	717.477	4.000	1336.000	.000	.682
	TB*SGOT*DB*SGPT	.392	215.767	4.000	1337.000	.000	.392
5-way Interactions	TB*DB*SGOT*SGPT*ALP	.683	574.015	5.000	1334.000	.000	.683

The combination of common attributes at different levels are 2-way relations, 3-way relations, 4-way relations and 5-way relations. The combination of attributes at 2-way relations are DB*ALP, DB*SGOT, DB*SGPT, SGOT*ALP, SGPT*ALP, SGOT*SGPT, TB*ALP, TB*DB, TB*SGOT and TB*SGPT. The combination of attributes at 3-way relations are DB*SGOT*ALP, DB*SGOT*SGPT, SGOT*SGPT*ALP, TB*DB*SGOT and TB*DB*SGPT. The combination of attributes at 4-way relations are DB*SGOT*SGPT*ALP, TB*DB*SGOT*ALP and TB*SGOT*DB*SGPT. The combination of attributes at 5-way Interaction is TB*DB*SGOT*SGPT*ALP. In this experimentation for the 95% of confidence level the equivalent significance level is 0.05. The Significant values are 0.000 for all the combinations of common attributes in multivariate analysis depicted in table 15 which is less than significant level i.e. 0.05. This indicates more significant difference exists between groups null hypothesis. For all the combinations that are DB*ALP, DB*SGOT, DB*SGPT, SGOT*ALP, SGPT*ALP, SGOT*SGPT, TB*ALP, TB*DB, TB*SGOT, TB*SGPT, DB*SGOT*ALP, DB*SGOT*SGPT, SGOT*SGPT*ALP, TB*DB*SGOT, TB*DB*SGPT, DB*SGOT*SGPT*ALP, TB*DB*SGOT*ALP, TB*SGOT*DB*SGPT and TB*DB*SGOT*SGPT*ALP they vary very much in population comparison.

Conclusions

The familiar attributes are ALP, DB, SGOT, SGPT and TB are considered from the three data sets for ANOVA and MANOVA. The significance level considered for the analysis is 0.05 with respect to confidence level is 95%. In this investigation ANOVA analysis shows more significant difference between groups due to rejection of null

hypothesis and it these groups designates more difference on ALP, DB, SGOT, SGPT and TB. Simultaneously in MANOVA analysis there exists more significant difference due to the null hypothesis based on significant value 0.000. The results indicate that there is more significant difference among three liver datasets that means there is a geographical effect on liver diseases. This may be due to food habits, alcoholic consumption, air pollution, life style etc. Then there is a need of localized modifications for the identification of liver diseases.

Future Scope

ANOVA and MANOVA analysis is also suggested for the various confidence levels like 99 % and 90 %. This statistical analysis may be applied for various regions of India i.e different states of India to investigate the geographical effect and to suggest the localized settings for the diagnosis of liver diseases.

Acknowledgments

I am thankful to UGC, New Delhi for granting financial support under UGC Minor Research Project scheme with reference no. MRP-6934/16 (SERO/UGC). I am also thankful to Prof. V. V. Nageswara Rao, Director of Aditya Institute of Technology and Management (A) for providing necessary facilities.

References

[1] Nibedan Panda Sir: Panda, N. and Majhi, S.K., 2020. Improved spotted hyena optimizer with space transformational search for training pi-sigma higher order neural network. Computational Intelligence, 36(1), pp.320-350.
 [2] Nibedan Panda Sir: Panda, N., Majhi, S.K., Singh, S. and Khanna, A., 2020. Oppositional spotted hyena optimizer with mutation

- operator for global optimization and application in training wavelet neural network. *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-14.
- [3] Nibedan Panda Sir: Panda, N. and Majhi, S.K., 2020. Improved salp swarm algorithm with space transformation search for training neural network. *Arabian Journal for Science and Engineering*, 45(4), pp.2743-2761.
- [4] Nibedan Panda Sir: Panda, N. and Majhi, S.K., Oppositional salp swarm algorithm with mutation operator for global optimization and application in training higher order neural networks. *Multimedia Tools and Applications*, pp.1-25.
- [5] Nibedan Panda Sir: Panda, N. and Majhi, S.K., 2019. How effective is spotted hyena optimizer for training multilayer perceptrons. *Int. J. Recent Technol. Eng*, pp.4915-4927.
- [6] Wong, R.J., Tran, T., Kaufman, H., Niles, J. and Gish, R., 2020. Geographic regions with high prevalence of nonalcoholic steatohepatitis-related hepatic fibrosis are also observed to demonstrate high prevalence of metabolic disease risk factors and low consumption of fruits and vegetables. *Clinical Nutrition Experimental*.
- [7] Xia, M., Sun, X., Zheng, L., Bi, Y., Li, Q., Sun, L., Di, F., Li, H., Zhu, D., Gao, Y. and Bao, Y., 2020. Regional difference in the susceptibility of non-alcoholic fatty liver disease in China. *BMJ Open Diabetes Research and Care*, 8(1), p.e001311.
- [8] Cannon, R.M., Davis, E.G., Goldberg, D.S., Lynch, R.J., Shah, M.B., Locke, J.E., McMasters, K.M. and Jones, C.M., 2020. Regional Variation in Appropriateness of Non-Hepatocellular Carcinoma Model for End-Stage Liver Disease Exception. *Journal of the American College of Surgeons*.
- [9] Jothimani, D., Venugopal, R., Abedin, M.F., Kaliamoorthy, I. and Rela, M., 2020. COVID-19 and Liver. *Journal of hepatology*.
- [10] Liao, H.Y., Hsu, J.C., Shia, B.C. and Lu, C.Y., 2020. PIN147 Geography Variations in Antibiotic Use RATE: Evidence from 8,026 Primary Clinics in Taiwan. *Value in Health*, 23, p.S568.
- [11] Khorraminezhad, L., Leclercq, M., Droit, A., Bilodeau, J.F. and Rudkowska, I., 2020. Statistical and Machine-Learning Analyses in Nutritional Genomics Studies. *Nutrients*, 12(10), p.3140.
- [12] Bendi, V.R. and Boddu, R.S.K., 2020. Performance Comparison of Classification Algorithms on Medical Datasets (No. 2322). *EasyChair*.
- [13] Percival, B., Gibson, M., Leenders, J., Wilson, P.B. and Grootveld, M., 2020. Univariate and Multivariate Statistical Approaches to the Analysis and Interpretation of NMR-based Metabolomics Datasets of Increasing Complexity.
- [14] Yousaf, M., Zahir, S., Riaz, M., Hussain, S.M. and Shah, K., 2020. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. *Chaos, Solitons & Fractals*, 138, p.109926.
- [15] Ustaoglu, F., Tepe, Y. and Taş, B., 2020. Assessment of stream quality and health risk in a subtropical Turkey river system: A combined approach using statistical analysis and water quality index. *Ecological Indicators*, 113, p.105815.
- [16] Petersen, M.W., Meyhoff, T.S., Helleberg, M., Kjær, M.B.N., Granholm, A., Hjortsø, C.J.S., Jensen, T.S., Møller, M.H., Hjortrup, P.B., Wetterslev, M. and Vesterlund, G.K., 2020. Low-dose hydrocortisone in patients with COVID-19 and severe hypoxia (COVID STEROID) trial—Protocol and statistical analysis plan. *Acta Anaesthesiologica Scandinavica*, 64(9), pp.1365-1375.
- [17] Gyimah, R.A.A., Gyamfi, C., Anornu, G.K., Karikari, A.Y. and Tsyawo, F.W., 2020. Multivariate statistical analysis of water quality of the Densu River, Ghana. *International Journal of River Basin Management*, pp.1-11.
- [18] Kattan, M.W. and Vickers, A.J., 2020. Statistical analysis and reporting guidelines for CHEST. *Chest*, 158(1), pp.S3-S11.
- [19] Ren, R., Zhou, H., Hu, Z., He, S. and Wang, X., 2019. Statistical analysis of fire accidents in Chinese highway tunnels 2000–2016. *Tunnelling and Underground Space Technology*, 83, pp.452-460.
- [20] Livadiotis, G., 2020. Statistical analysis of the impact of environmental temperature on the exponential growth rate of cases infected by COVID-19. *PLoS one*, 15(5), p.e0233875.
- [21] Verma, A. and Ranga, V., 2018. Statistical analysis of CIDDS-001 dataset for network

- intrusion detection systems using distance-based machine learning. *Procedia Computer Science*, 125, pp.709-716.
- [22] Zajkowska, M., Gacuta, E., Kozłowska, S., Lubowicka, E., Głażewska, E.K., Chrostek, L., Szmitkowski, M., Pawłowski, P., Zbucka-Krętowska, M. and Ławicki, S., 2019. Diagnostic power of VEGF, MMP-9 and TIMP-1 in patients with breast cancer. A multivariate statistical analysis with ROC curve. *Advances in medical sciences*, 64(1), pp.1-8.
- [23] Mustaqeem, A., Anwar, S.M., Khan, A.R. and Majid, M., 2017. A statistical analysis based recommender model for heart disease patients. *International journal of medical informatics*, 108, pp.134-145.
- [24] Zhao, S., Gao, D., Zhuang, Z., Chong, M.K., Cai, Y., Ran, J., Cao, P., Wang, K., Lou, Y., Wang, W. and Yang, L., 2020. Estimating the serial interval of the novel coronavirus disease (COVID-19): A statistical analysis using the public data in Hong Kong from January 16 to February 15, 2020.
- [25] Lara-Cabrera, R., Gonzalez-Pardo, A. and Camacho, D., 2019. Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. *Future Generation Computer Systems*, 93, pp.971-978.
- [26] Barbulescu, A. and Nazzal, Y., 2020. Statistical analysis of dust storms in the United Arab Emirates. *Atmospheric Research*, 231, p.104669.
- [27] Sidorov, O.V., Kozub, L.V., Gofenberg, A.V. and Osintseva, N.V., 2018. Organization and Carrying out the Educational Experiment and Statistical Analysis of Its Results in IHL. *European Journal of Contemporary Education*, 7(1), pp.177-189.
- [28] Nordhausen, K. and Oja, H., 2018. Independent component analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5), p.e1440.

Technical Report submitted to UGC on

**Analysis of Geographical effect of various regions of
INDIA on Liver disease using Data Mining Techniques**

**UGC Minor Research Project sanction F.No.MRP-6934/16
(UGC-SERO)-dated 02/08/2017**

Dr. BENDI VENKATA RAMANA

Principal Investigator

Professor & HOD, Dept. of Information Technology



Department of Information Technology

Aditya Institute of Technology and Management (A)

K. Kotturu, Tekkali, Srikakulam-532201, A.P., India.

Approved by AICTE, Permanently Affiliated to JNTUK Kakinada, Accredited by NBA,

Accredited by NAAC (UGC) with A+, Recognized by UGC u/s 2(f) & 12 (B),

Recognized as SIRO by DSIR

Summary

Liver disease is the tenth most common cause of death in India as per the World Health Organization. Sedentary lifestyle, fatty food, uncontrolled blood sugar, obesity, smoking and high alcohol intake is leading Indians towards higher incidence of liver disease. Earlier detection can reduce the death rates and it also helps the doctors to give the proper treatment to the patients. The liver disease datasets are analyzed by using Data Mining techniques and Statistical techniques for the accurate diagnosis of liver disease and to know the impact of geographical effect on Liver diseases respectively.

The datasets were collected and annotated from Visakhapatnam, Vijayawada and Tirupathi based on the major geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema respectively. The three datasets are named Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. Visakhapatnam dataset contains 12 attributes and has 499 samples. Vijayawada dataset contains 12 attributes and has 600 samples. Tirupathi dataset contains 7 attributes and has 243 samples.

Data mining techniques plays a major role in the classification of Diseases. The selected Classification Algorithm Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron were considered for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time in classifying liver patients dataset. Classification performance is very high in Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets, where as Classification performance is very high in Random Forest classification algorithm for

Vijayawada dataset. Building time is more for MLP in Vijayawada dataset. This study motivated towards for the development of Liver Diagnosis App using Decision Tree classifier for the benefit of the liver patients and doctors.

Statistical Analysis plays a significant role in population comparison to investigate the geographical effect on liver diseases. In this study the common attributes ALP, DB, SGOT, SGPT and TB were considered from the three datasets for the population comparison. Three data sets were evaluated using analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). The significance level considered for the statistical analysis is 0.05, the corresponding confidence level is 95%.

The Significant values in the ANOVA and MANOVA analysis indicates there is more significant difference among three liver datasets that means there is a geographical effect on liver diseases. This may be due to food habits, alcoholic consumption, air pollution, life style etc. Then there is a need of localized modifications for the identification of liver diseases.

Statistical analysis is also suggested for the various confidence levels like 99 % and 90 % and may be applied for various regions of India i.e different states of India to investigate the geographical effect and to suggest the localized settings for the diagnosis of liver diseases.

CERTIFICATE

This is to certify that the project work titled “Analysis of Geographical effect of various regions of INDIA on Liver disease using Data Mining Techniques” is carried out by me and was not submitted for partial/full financial assistance to any other funding agency.

Dr. Bendi Venkata Ramana
Principal Investigator

ACKNOWLEDGEMENTS

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals. I am highly indebted to the University Grant Commission-SERO, Hyderabad for considering my proposal and providing financial assistance.

I am highly indebted to Prof.V.V. Nageswra Rao, Director of Aditya Institute of Technology and Management (A) and Dr. A. Srinivasa Rao, Principal for the facilities provided to accomplish this project. I am obliged to Dr. K.B.Madhu Sahu, Director R&D and R&D team for their encouragement.

I am thankful to the Management, Aditya Institute of Technology and Management for providing necessary facilities in the department of Information Technology. I am also thankful to all the senior members of the department for their constructive criticism throughout the project.

My heartfelt thanks are due to all my colleagues in the department for their keen interest and great support throughout the project.

I express my gratitude to non teaching staff of our department for their cooperation throughout the project.

Finally, I express my heartfelt thanks to all of my family members who helped me in successful completion of this project.

Dr. Bendi Venkata Ramana
Principal Investigator

CONTENTS

Chapter-1		Page No : 1-4	
S.No	Introduction	From	To
1.1	Motivation	2	2
1.2	Data Sets	2	3
1.3	Data Mining Techniques	4	4
1.4	Statistical Analysis	4	4

Chapter-2		Page No : 5-11	
S.No	Literature Survey	From	To
2.1	Introduction	6	6
2.2	Review on Classification Algorithms	6	8
2.3	Review on Statistical Analysis	8	9
2.4	Summary	9	9
2.5	Origin of the Research Problem	10	10
2.6	Social Benefit	10	10
2.7	Objectives	10	11
2.8	Methodology	11	11

Chapter-3		Page No : 12-30	
S.No	Performance Analysis of Liver Disease Diagnosis using Data Mining Techniques	From	To
3.1	Introduction	13	13
3.2	Data Mining Techniques	13	13
3.3	Supervised Learning	14	14
3.3.1	Naive Bayes Algorithm	14	15
3.3.2	Decision Tree Algorithm	15	15
3.3.3	Random Forest Algorithm	15	16
3.3.4	Support Vector Machines	16	16
3.3.5	Neural Network Algorithm	17	17
3.4	Performance Evaluation	18	21
3.5	Results and Discussion	22	30
3.6	Conclusions	30	30
3.7	Future Scope	30	30

Chapter-4		Page No : 31-44	
S.No	Analysis of Geographical effect of various regions on Liver disease	From	To
4.1	Introduction	32	32
4.2	Statistical Analysis	32	33
4.2.1	One way analysis of variance (ANOVA):	33	34
4.2.2	Multivariate Analysis of variance (MANOVA):	34	34
4.3	Results and Discussion	35	44
4.4	Conclusions	44	44
4.5	Future Scope	44	44

Chapter-5		Page No : 45-46	
S.No	Conclusions	From	To
5.1	Conclusions	46	46
5.2	Future Scope	46	46
	References	47	50
	Appendix		

CHAPTER 1

INTRODUCTION

1.1: Motivation

Liver disease is the tenth most common cause of death in India as per the World Health Organization. Sedentary lifestyle, fatty food, uncontrolled blood sugar, obesity, smoking and high alcohol intake is leading Indians towards higher incidence of fatty liver disease, that is "Non Alcoholic Fatty Liver Disease". Non-alcoholic fatty liver disease is now one of the most common causes of chronic liver disease. Earlier detection can reduce the death rates and it also helps the doctors to give the proper treatment to the patients. To make the process simple Data Mining techniques need to be utilized for accurate diagnosis of liver diseases.

1.2: Data Sets

The datasets were taken from Visakhapatnam, Vijayawada and Tirupathi based on the major geographical regions of Andhra Pradesh that are North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema respectively. These datasets are analyzed by using Data Mining techniques for the accurate diagnosis of liver disease and to know the impact of geographical variables such as food habits, behaviors, environment etc on Liver Function Tests (LFT). The three datasets are named Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. Visakhapatnam dataset contains 12 attributes and has 499 samples. Vijayawada dataset contains 12 attributes and has 600 samples. Tirupathi dataset contains 7 attributes and has 243 samples. The description of datasets are given table1. The list and type of attributes of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset are represented table 2, table 3 and table 4 respectively.

Table 1. Description of datasets from different regions of Andhra Pradesh

Name of the Dataset	Number of Attributes	Number of Samples	Number of Classes
Tirupathi Dataset	12	499	2
Vijayawada Dataset	12	600	2
Visakhapatnam Dataset	7	243	2

Table 2. Visakhapatnam Liver Dataset Attributes and Type

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
SGOT	Integer
SGPT	Integer
ALP	Integer

Table 3. Vijayawada Liver Dataset Attributes and Type

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
AST (SGOT)	Integer
ALT (SGPT)	Integer
ALP	Integer
IB	Real number
SP (TP)	Real number
SA (Albumin)	Real number
SG (Globulins)	Real number
A/G RATIO	Real number

Table 4. Tirupathi Liver Dataset Attributes and Type

Attribute	Type
AGE	Real number
GENDER	Categorical
TB	Real number
DB	Real number
SGOT	Integer
SGPT	Integer
ALP	Integer
IB	Real number
TP	Real number
Albumin	Real number
Globulins	Real number
A/G RATIO	Real number

1.3: Data Mining Techniques

Data Mining are widely used in the health care industry for predicting the diseases from the datasets. Data mining algorithms are categorized into unsupervised learning (Classification) and supervised learning (clustering). The liver datasets considered are suitable for supervised learning. Data Mining techniques considered in this research are classification techniques that are Naive Bayes, Random Forest, Decision Tree and Multilayer perceptron (MLP). These classification techniques are used for the accurate diagnosis of the liver disease.

1.4: Statistical Analysis

Statistical analysis involves the test of the relationship between two statistical data sets. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis of no relationship between two data sets. Rejecting the null hypothesis is done using statistical tests. One-way Analysis of Variance (ANOVA) and Multivariate Analysis of Variance (MANOVA) are popularly used statistical analysis techniques for evaluating the significant difference between populations. ANOVA is used to test the significant difference in a single dependent variable among two or more groups formed by a single independent or classification variable, whereas MANOVA is used to test the significant difference in more than one dependent variable and several independent variables.

CHAPTER 2

LITERATURE SURVEY

2.1: Introduction

The classification algorithms Naive Bayes, Decision Tree, Random Forest and Multilayer Perceptron are applied on three datasets named as Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset for the identification of best classification algorithm for the disease diagnosis. ANOVA and MANOVA statistical techniques are considered for the population comparison between datasets to know the geographical effect on liver datasets. A brief review on comparison algorithms is presented in 2.2. A review on Statistical analysis is presented in 2.3.

2.2: Review on Classification Algorithms

Lung-Cheng Huang [1] reported that Naive Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset.

Rafael Berlanga et al. [2] developed an integrated health care platform for European pediatrics and decision support tools to access personalized health information and introduces both the integrated data model in the Health-e-Child project.

Kemal Polat et al. [3] applied Fuzzy AIRS (Artificial Immune Recognition System) classification algorithm in the diagnosis of Breast Cancer and Liver Disorders.

Humar Kahramanli and Novruz Allahverdi [4] presented ANN algorithm and applied to BUPA Liver Disorders dataset for extracting comprehensible classification rules for diagnosis of liver disorders

Paul R. Harper [5] considered a selected classification algorithms and evaluated their relative performances and practical usefulness on different types of health care datasets and reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed.

Rong Ho Lin [6] proposed an intelligent model for the diagnosis of liver diseases which integrates classification and regression tree (CART) and case-based reasoning (CBR) techniques. The results indicate that the CART rate of accuracy is 92.94% and CBR diagnostic accuracy rate is 90.00%.

Bendi Venkataramana et al [7] considered selected classification algorithms for the classification of liver patient datasets based on four criteria: Accuracy, Precision, Sensitivity and Specificity.

Y. Rakhshani Fatmehsari and F. Bahrami [8] presented a new algorithm to analyze gait patterns of subjects with Parkinson's disease (PD) under two conditions of deep brain stimulation (DBS) off and on.

Mohd Fauzi Othman and Mohd Ariffanan Mohd Basri [9] proposed 'Probabilistic Neural Network' with image and data processing techniques was implemented to an automated brain tumor classification.

Li Bin Ren et al. [10] proposed a Bayesian based BTS classification algorithm (b-BTS) and compared with other multi-class SVM, Binary Tree of SVM (BTS) takes a good advantage of lower time consuming.

Ying Li and Bo Cheng [11] applied an improved KNN algorithm in the classification of image objects obtained by segmentation. Their experiment shows that in the same training set and testing set, the improved KNN algorithm can achieve higher accuracy in the classification of high resolution remote sensing image.

Giorgia Macchiavello et al. [12] addressed the problem of identifying snow-covered areas without training information by combining a decision tree approach with a Bayesian thresholding method.

A. Mathur and G. M. Foody [13] demonstrated Support vector machines (SVMs) have considerable potential for supervised classification analyses, but their binary nature has been a constraint on their use in remote sensing.

Parviz Zeaiean Firouzabadi [14] proposed different supervised classification algorithms were applied to estimate crop acreage using similar training sites.

Yu-guo Wang and Hua-peng Li [15] proposed a method for sample purification based on statistical analysis theory which could purify training samples for improved wetlands remote sensing classification based on ANN.

Zhou Faguo et al. [16] proposed the short text classification based on statistics and rules on the foundation of several common used classic texts classification algorithms, mainly according to the major feature extraction methods. Experiments show that this algorithm has better performance than other algorithms.

Zheng-Tao Yu et al. [17] presented a domain text classification model. This model uses the support vector machine learning algorithm and domain knowledge. Their results proved that domain knowledge relations have a good influence on the domain text classification.

Ayse Cufoglu et al. [18] compared four different classification algorithms which are; Naïve Bayesian (NB), Instance-Based Learner (IB1), Bayesian networks (BN) and Lazy Learning of Bayesian Rules (LBR) classifiers. Their simulation results show that, the NBTree has the highest classification accuracy performance with the lowest error rate.

Noor Ezan Abdullah et al. [19] presented various supervised ANN models for plaque classification using RGB indices. These models are designed and implemented Levenberg Marquardt feed forward, Radial Basis Function algorithm and back propagation.

2.3: Review on Statistical Analysis

B.Surendiran et al. [20] proposed an Univariate Analysis Of Variance (ANOVA) Discriminate Analysis (DA) classifier for classifying the masses present in mammogram. Experimental results shows that the proposed

method reaches high classification accuracy in compared to existing algorithms.

Mireille Tohm'e et al. [21] proposed an alternative to usual multiclass multivariate group comparison tests such as MANOVA or Wilcoxon tests to compare drugs in high dimensional spaces.

Bendi Venkataramana et al [22] proposed ANOVA, MANOVA analysis on liver patient datasets of INDIA and USA and observed that liver patients of both the countries are having significant difference which is the reason for difference in classifiers performance. Results of this study are very important for the development of automatic medical diagnosis system and the need for its localization settings based on the geographical region.

Z. A. Dastgheib et al. [23] Introduced a novel method based on analysis of dynamic response of vestibular system for diagnosis of Parkinson's disease (PD). Multivariate analysis of variance (MANOVA) was used to select pairs of features showing the most significant differences between the groups.

S. Dimitrova [24] employed MANOVA to check the significance of the influence planetary geomagnetic activity, gender and medication. Results obtained suggest that females and persons on a medication are more sensitive to geomagnetic activity increment in comparison with respectively males and persons not taking medicaments.

2.4: Summary

This survey is made on classification algorithms and statistical analysis. Review on classification algorithms indicates that no single classifier is suitable for all datasets. It is necessary to identify suitable classifier for the accurate diagnosis of various diseases and also evaluates the performance of the classifier with various performance evaluators. Review on statistical analysis motivates to identify the geographical effect on liver disease.

2.5: Origin of the Research Problem:

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patients' survival rate. Liver disease can be diagnosed by analysing the levels of enzymes in the blood. Moreover, now a day's mobile devices are extensively used for monitoring humans' body conditions. Here also, automatic classification algorithms are needed. With the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as gastroenterologist.

Recent research studies on liver diagnosis indicated difference in classification accuracy of various classifiers with different data sets. In order to envisage the reason for this difference, we propose to analyze the liver patient's populations of various regions of India.

2.6: Social Benefit:

This project is very much useful to know the status of the liver at an early stage and also identify the geographical impact on liver. Results of this study are very important for the development of automatic medical diagnosis system and the need for its localization settings based on the geographical region.

2.7: Objectives:

The major objectives of this project are as follows:

To diagnose the reasons for LFT using various data mining techniques.

To know the impact the food habits on liver disease of different regions of Andhra Pradesh

To analyze the impact of different surroundings and behavioral habits of people on liver diseases.

To study the comparative analysis of liver patients of different regions of Andhra Pradesh.

2.8: Methodology:

In this project liver patient's dataset from different geographical regions of Andhra Pradesh will be collected and used for the analysis to know the impact of geographical variables such as food habits, behaviors, environment etc on LFT.

The present study will be conducted in Andhra Pradesh. The state can be divided into 3 major geographical regions namely North Coastal Andhra Pradesh, Central Andhra Pradesh and Rayalaseema. One growth center from each region will be selected, where hospital facilities are available to diagnose LFT. The geographical regions and corresponding growth centers are North Coastal Andhra Pradesh, Central Andhra Pradesh, Rayalaseema and Visakhapatnam, Vijayawada, Titupathi respectively.

CHAPTER 3

Performance Analysis of Liver Disease Diagnosis using Data Mining Techniques

3.1: Introduction

Data mining techniques play a major role in the classification of Diseases. It involves data mining algorithms to analyze medical data. Day by day, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in India. Liver patient datasets are investigated using selected classification algorithms. The datasets considered are Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset based on geographical region. The selected classification algorithms considered are Naive Bayes algorithm, Decision Tree Algorithm, Random Forest Algorithm and Neural Network Algorithm. Data Mining techniques are also used popularly for automated liver disease diagnosis. Liver disease are classified based on the liver function Tests (LFT).

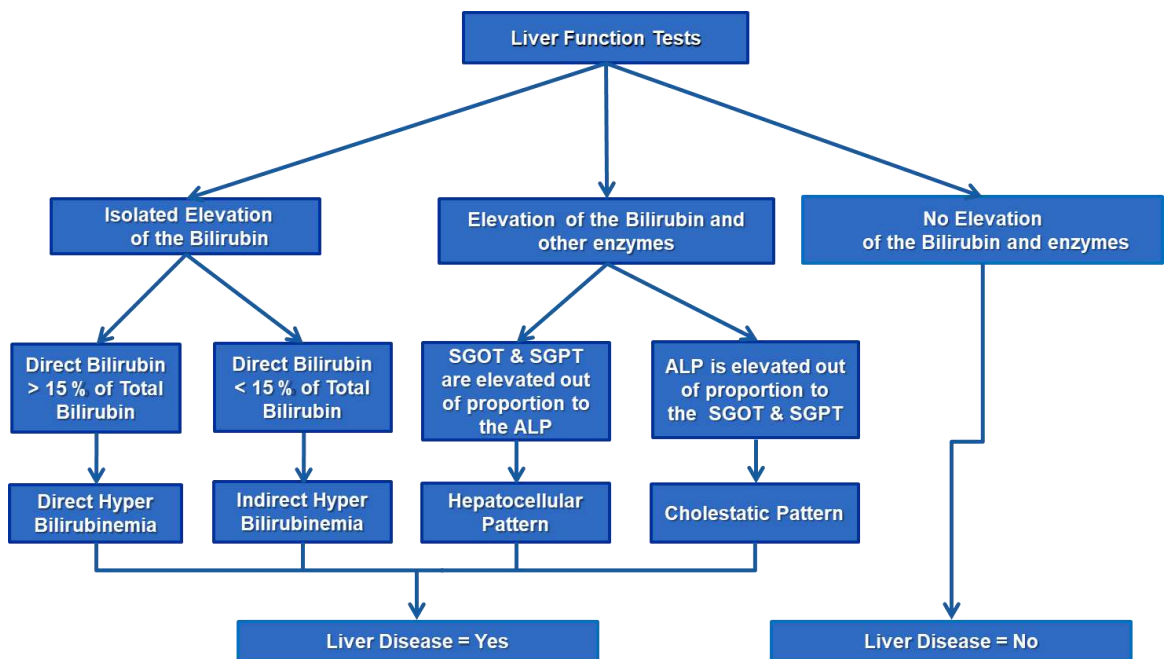


Fig. 1. Process of Liver Disease Classification

3.2: Data Mining Techniques

Data Mining is a multidisciplinary field that uses machine learning, statistics, AI and database technology. The popular data mining techniques are Classification, Clustering, Regression, Association Rules, Outer detection, Sequential Patterns and Prediction.

3.3: Supervised Learning

A supervised learning algorithm learns from labeled training data and predicts the class of unforeseen data. Supervised learning uses classification and regression techniques to develop predictive models. Classification algorithms are popularly used in various medical applications. Data classification is a two step process in which first step is the training phase where the classifier algorithm builds classifier with the training set of tuples and the second step is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples.

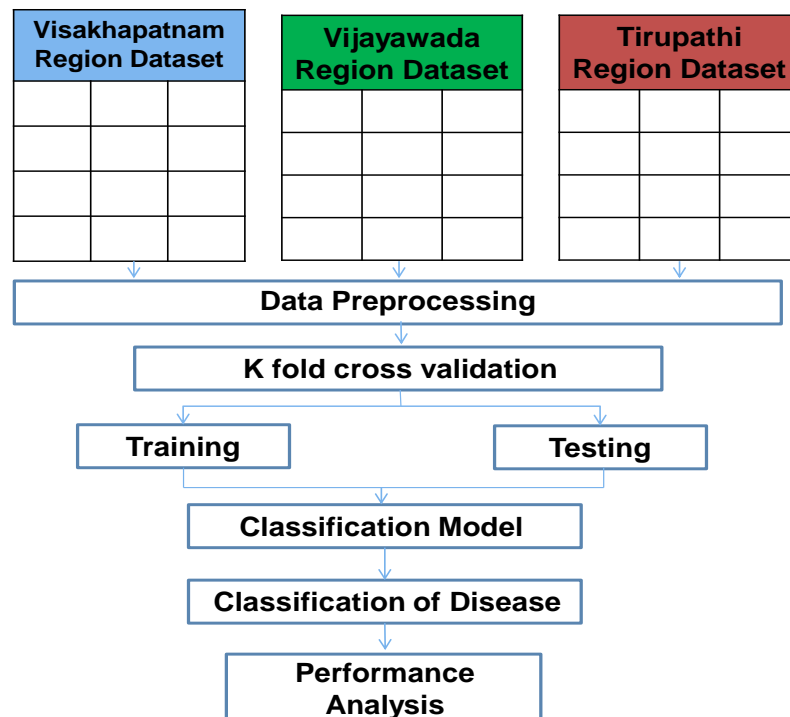


Fig. 2. Process of performance analysis

3.3.1: Naive Bayes Algorithm

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, independence between attributes of data points. A naive Bayes classifier uses probability theory to classify data.

Bayes' Theorem is stated as: $P(h | d) = (P(d | h) * P(h)) / P(d)$

Where $P(h|d)$ is the probability of hypothesis h given the data d . This is called the posterior probability.

$P(d|h)$ is the probability of data d given that the hypothesis h was true.

$P(h)$ is the probability of hypothesis h being true (regardless of the data).

This is called the prior probability of h .

$P(d)$ is the probability of the data (regardless of the hypothesis).

3.3.2: Decision Tree Algorithm

Decision trees are the most powerful algorithms that falls under the category of supervised algorithms. They can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions

Input: D dataset – features with a target class

for \forall features do

for Each sample do

Execute the Decision Tree algorithm

end for

Identify the feature space f_1, f_2, \dots, f_x of dataset UCI.

end for

Obtain the total number of leaf nodes $l_1, l_2, l_3, \dots, l_n$ with its constraints

Split the dataset D into $d_1, d_2, d_3, \dots, d_n$ based on the leaf node constraints.

Output: Partition datasets d_1, d_2, d_3, \dots

3.3.3: Random Forest Algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression.

- First, start with the selection of random samples from a given dataset.

- Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- In this step, voting will be performed for every predicted result.
- At last, select the most voted prediction result as the final prediction result.

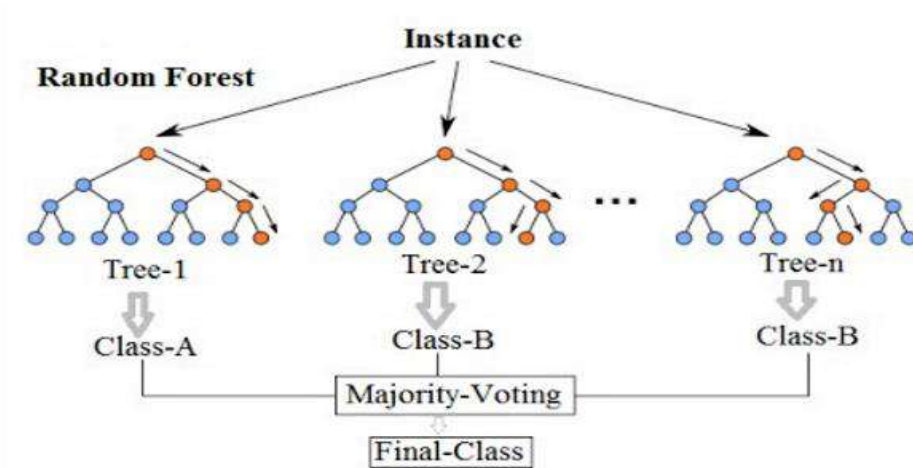


Fig. 3. Random Forest Algorithm

3.3.4: Support Vector Machines

Support vector machines (SVMs) are powerful and flexible supervised machine learning algorithms which are used both for classification and regression. The main objective of SVM is to find the optimal hyperplane which linearly separates the data points in two component by maximizing the margin.

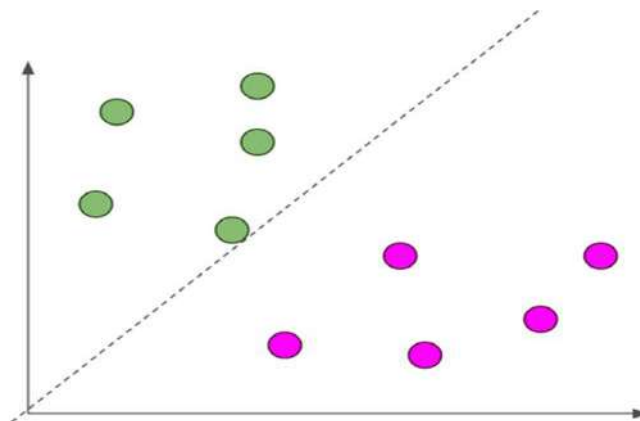


Fig. 4. Support Vector Machines

3.3.5: Neural Network Algorithm

Feedforward supervised neural networks were among the first and most successful supervised learning algorithm. They are also called deep networks, multi-layer Perceptron (MLP), or simply neural networks and the vanilla architecture with a single hidden layer is illustrated. Each Neuron is associated with other neuron with some weight, The network processes the input upward activating neurons as it goes to finally produce an output value. This is called a forward pass on the network. A multilayer perceptron (MLP) is a class of feed forward artificial neural network (ANN). MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. Multilayer perceptron is the most frequently used type of neural network. Its architecture is called feed forward, the signals in the network are transmitted in one direction, input to output.

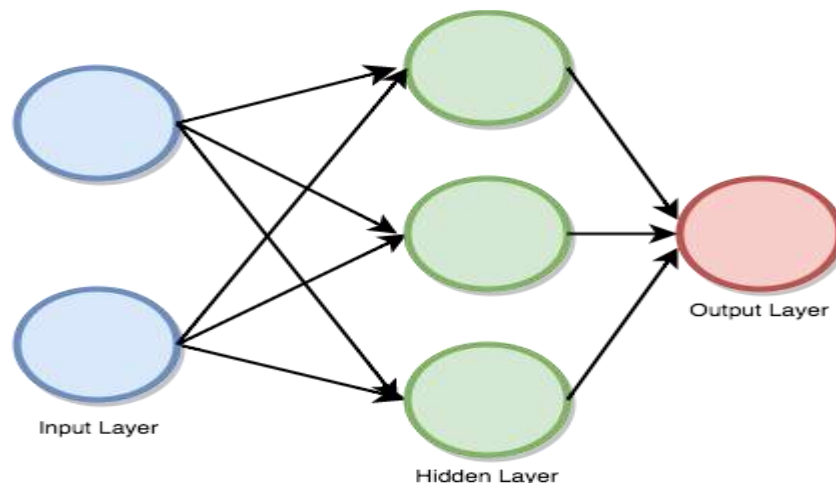


Fig. 5. Feed Forward Artificial Neural Network

3.4: Performance Evaluation

Evaluation of the performance of a classification model is based on the counts of (testing) objects correctly and incorrectly predicted. Confusion Matrix describes the performance of a classification model.

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. Consider a two-class prediction problem (binary classification), in which the outcomes are labeled either as positive (p) or negative (n). There are four possible outcomes from a binary classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP); however if the actual value is n then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are n, and false negative (FN) is when the prediction outcome is n while the actual value is p.

		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FP
	Negative	FN	TN

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Precision: precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Sensitivity: Sensitivity is also referred as True positive rate i.e the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

Specificity: Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified.

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

False Positive Rate

False Positive Rate (FPR) or “Fall-Out” is the proportion of negative cases incorrectly identified as positive cases in the data. FPR is 1-Specificity.

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$$

F-Measure

The F-Score is the Harmonic mean of precision and recall. The F-Score (F-Measure) is a single measure of classification procedure’s usefulness. The F-Score considers both precision and recall of the procedure to compute score. The higher the F-score, the better the predictive power of the classification procedure. A score of 1 means the classification procedure is perfect. The lowest possible F-score is 0.

$$\text{F - Measure} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. An ROC curve plots True Positive Rate (TPR) vs. False Positive Rate (FPR) at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

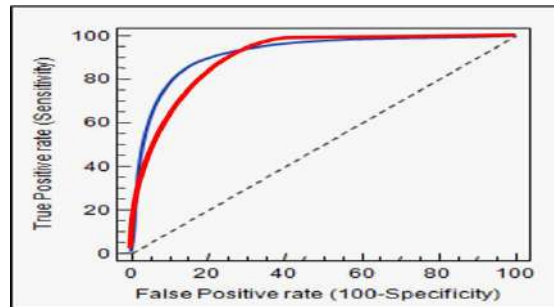


Fig. 6. Roc Curve

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Mean Absolute Error

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\mathbf{MAE} = \mathbf{1/n} \sum \mathbf{|y_j - y^{\wedge}j|}$$

Root Mean Squared Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far liver dataset from the regression line data points. RMSE is a measure of spread out these residuals is liver dataset. It's the square root of the average of squared differences between prediction and actual observation.

$$\mathbf{RMSE} = \sqrt{\mathbf{(f-o) 2}}$$

f = forecasts (expected values or unknown results),

o = observed values (known results).

Root Relative Squared Error

The Root Relative Squared Error (RRSE) is the Root Mean Squared Error (RMSE) divided by the Root Mean Prior Squared Error (RMPSE). The root relative squared error is relative to what it would have been if a simple liver disease predictor and just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

Kappa Statistics

Kappa statistics provides data to recognize the level to which physician persuade with each other further than what you might be expecting to see based on likelihood alone. It indicates the proportion of agreement beyond that expected by chance. The kappa coefficient provides valuable information on the reliability of diagnostic and other examination procedures.

$$\mathbf{K = \text{Observed agreement} - \text{chance agreement} / (1 - \text{chance agreement})}$$

Building Time

Building Time is the time required for the classifier model construction. A classification model tries to draw some conclusion from the input values given for training. It is simply the training time required for the classifier.

3.5: Results and Discussion

In this study the selected classification algorithms were considered for the evaluation of the performance of classifier in the diagnosis of liver disease. The classification algorithms considered in this study are Naive Bayes algorithm, Decision Tree algorithm, Random Forest algorithm, Support Vector Machines (SVM) and Multi Layer Perceptron (MLP).

The three liver datasets from various regions of Andhra Pradesh were considered for the evaluation of classification algorithms based on the various performance evaluaters. The performance evaluaters are Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time.

In this experimentation 10 fold cross validation was considered for each classification algorithm. That means each dataset was divided into ten parts out of which nine parts were used as training set and the remaining part is used as testing set. Repeating these ten folds ensures that each part is used for training and testing thus minimizing the sample bias.

The performance measures and error measures are evaluated for the Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron on Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset. These measures are depicted in table 5.

Performance comparison of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset for the Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron are represented in Fig. 7, Fig. 9 and Fig. 11 respectively.

Error comparison of Visakhapatnam dataset, Vijayawada dataset and Tirupathi dataset for the Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron are represented in Fig. 8, Fig. 10 and Fig. 12 respectively.

Table 5. Performance evaluation of selected classifiers for the three liver data sets

Dataset \ Algorithm		Naive Bayes	Decision Tree	Random Forest	SVM	MLP
Visakhapatnam Data set	Accuracy	93.4156	100	100	83.1276	96.7078
	Precision	0.934	1.000	1.000	0.838	0.967
	Sensitivity	0.934	1.000	1.000	0.831	0.967
	Specificity	0.9402	1.000	1.000	0.8246	0.9722
	F-Measure	0.933	1.000	1.000	0.810	0.967
	ROC-Area	0.951	1.000	1.000	0.704	0.978
	FPR	0.131	0.000	0.000	0.423	0.061
	MAE	0.0641	0	0.0127	0.1687	0.0526
	RMSE	0.2473	0	0.0519	0.4108	0.1747
	RRSE	55.857	0	11.7123	92.7776	39.4544
	Kappa Statistic	0.8269	1	1	0.4867	0.9152
	Building Time (Sec)	0	0	0.04	0.06	0.24
Vijayawada Data set	Accuracy	80.8333	97.6667	98.5	77.1667	84.3333
	Precision	0.800	0.977	0.986	0.786	0.839
	Sensitivity	0.808	0.977	0.985	0.772	0.843
	Specificity	0.8146	0.9866	0.9977	0.7697	0.8896
	F-Measure	0.779	0.977	0.985	0.697	0.841
	ROC-Area	0.790	0.964	0.993	0.547	0.865
	FPR	0.508	0.035	0.009	0.678	0.294
	MAE	0.198	0.0246	0.0438	0.2283	0.1798
	RMSE	0.4169	0.1467	0.1226	0.4778	0.3506
	RRSE	96.5015	33.9636	28.3694	110.5979	81.1429
	Kappa Statistic	0.3689	0.9378	0.9604	0.1332	0.5667
	Building Time (Sec)	0	0.03	0.23	0.08	1.02
Tirupathi Data set	Accuracy	98.5972	100	99.7996	99.5992	99.5992
	Precision	0.986	1.000	0.998	0.996	0.996
	Sensitivity	0.986	1.000	0.998	0.996	0.996
	Specificity	0.9952	1.000	0.9953	0.9907	0.9953
	F-Measure	0.986	1.000	0.998	0.996	0.996
	ROC-Area	1.000	1.000	0.999	0.996	0.995
	FPR	0.018	0.000	0.002	0.003	0.004
	MAE	0.0365	0	0.0146	0.004	0.006
	RMSE	0.1118	0	0.0517	0.0633	0.0641
	RRSE	22.5796	0	10.4514	12.7914	12.943
	Kappa Statistic	0.9713	1	0.9959	0.9918	0.9918
	Building Time (Sec)	0.01	0.03	0.24	0.15	0.9

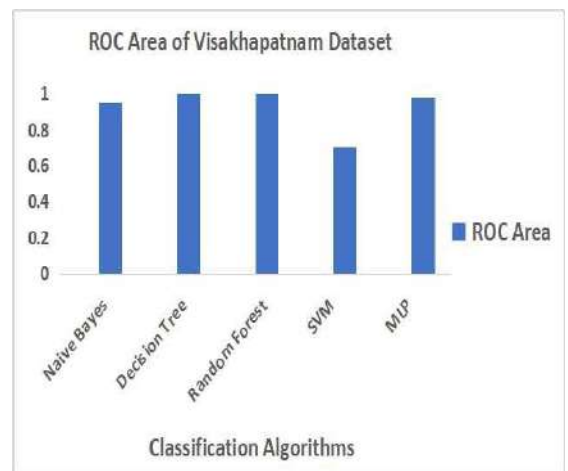
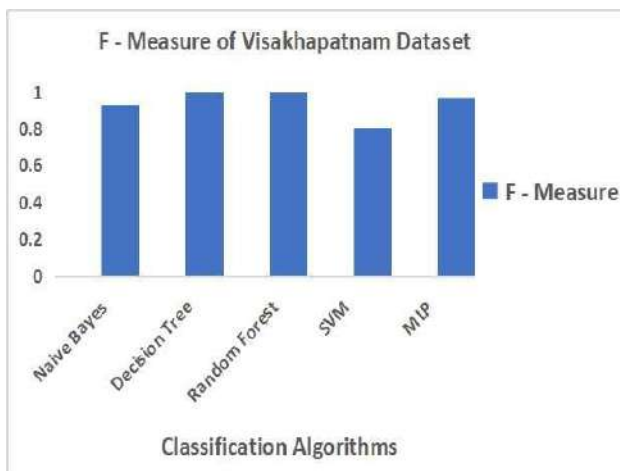
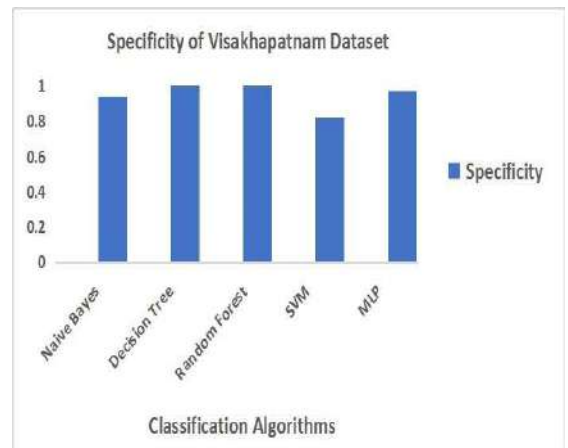
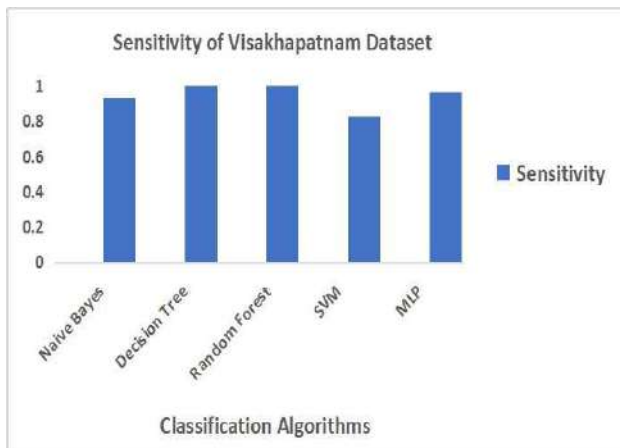
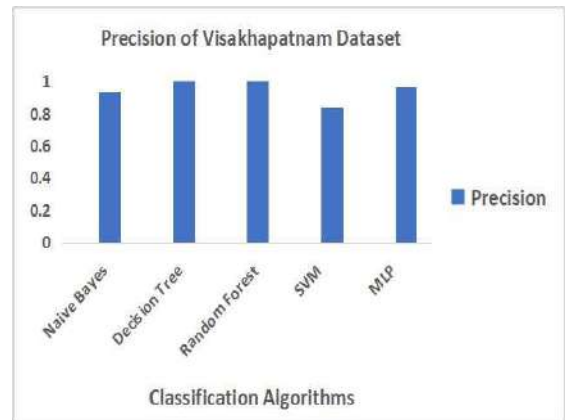
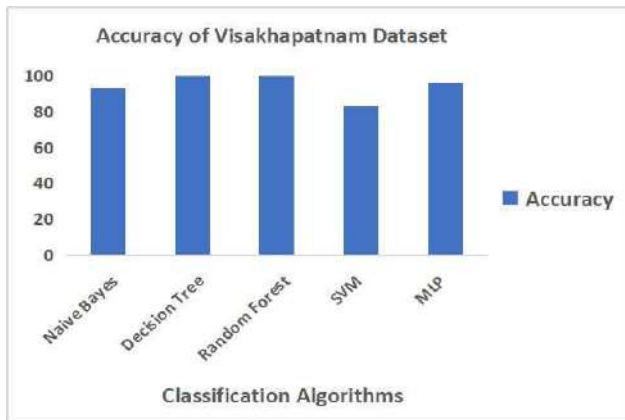


Fig.7. Performance comparison of Visakhapatnam dataset

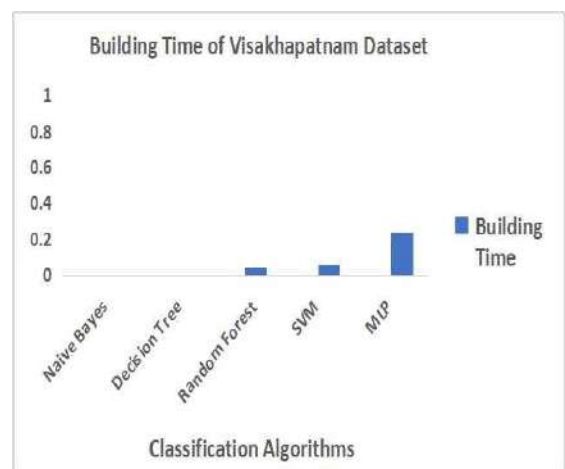
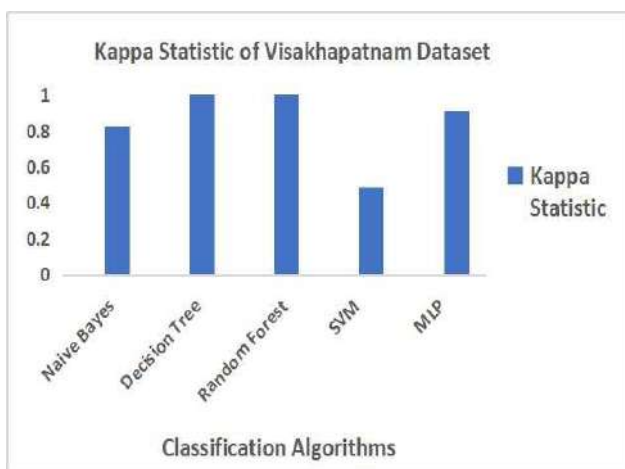
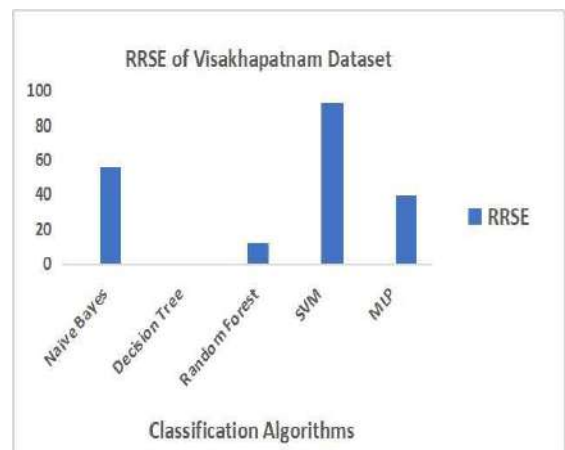
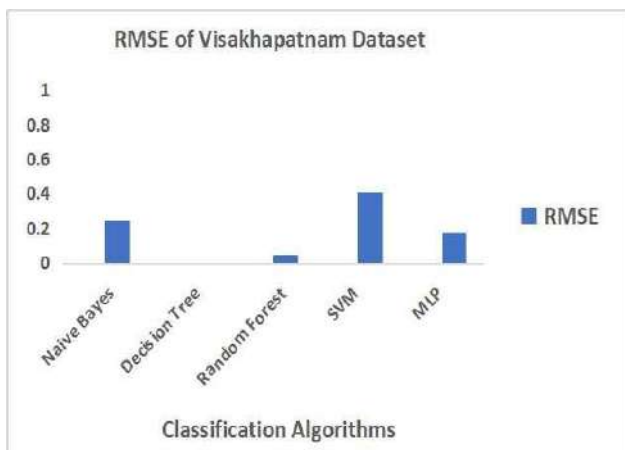
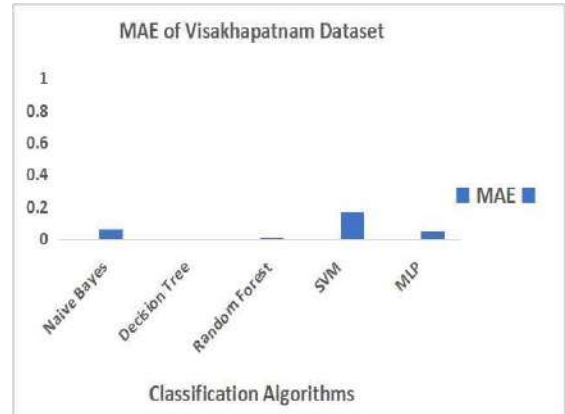
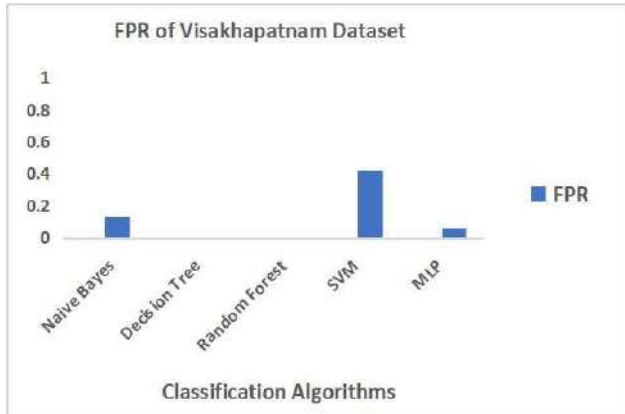


Fig. 8. Error comparison of Visakhapatnam dataset

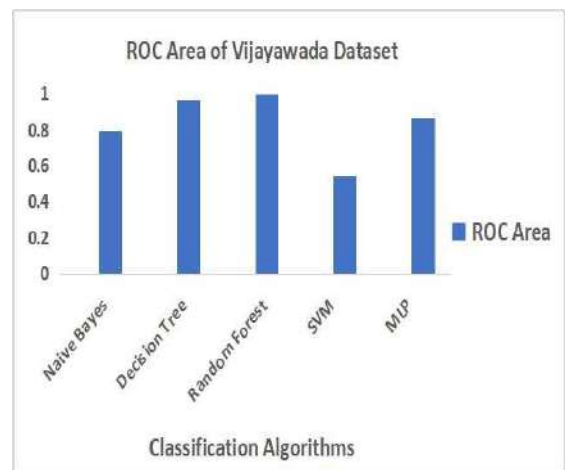
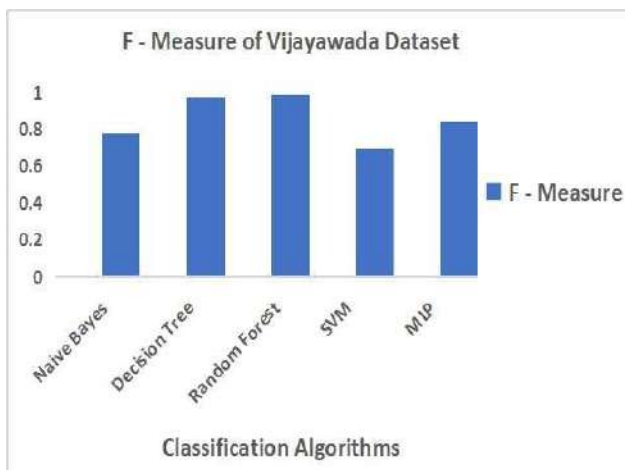
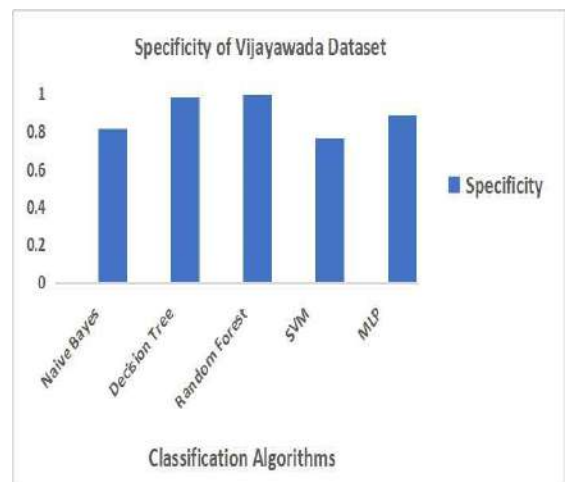
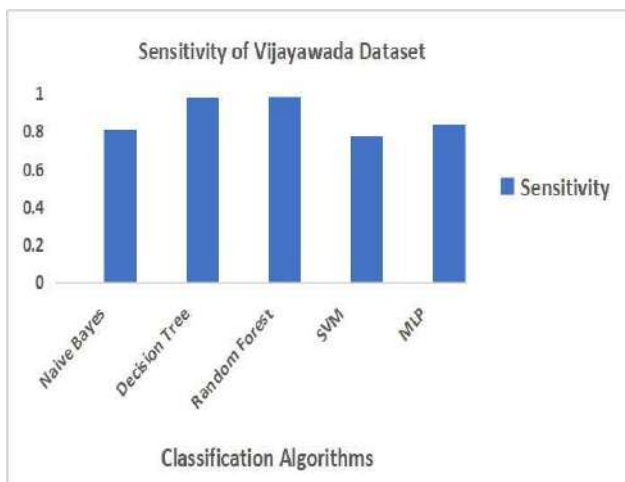
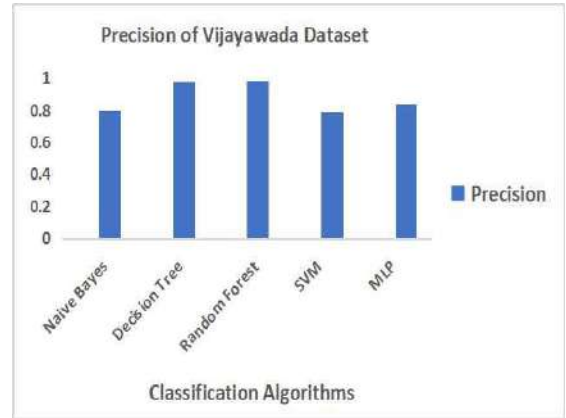
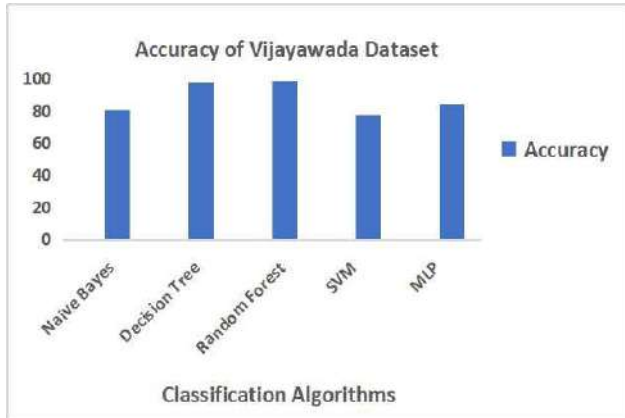


Fig. 9. Performance comparison of Vijayawada dataset

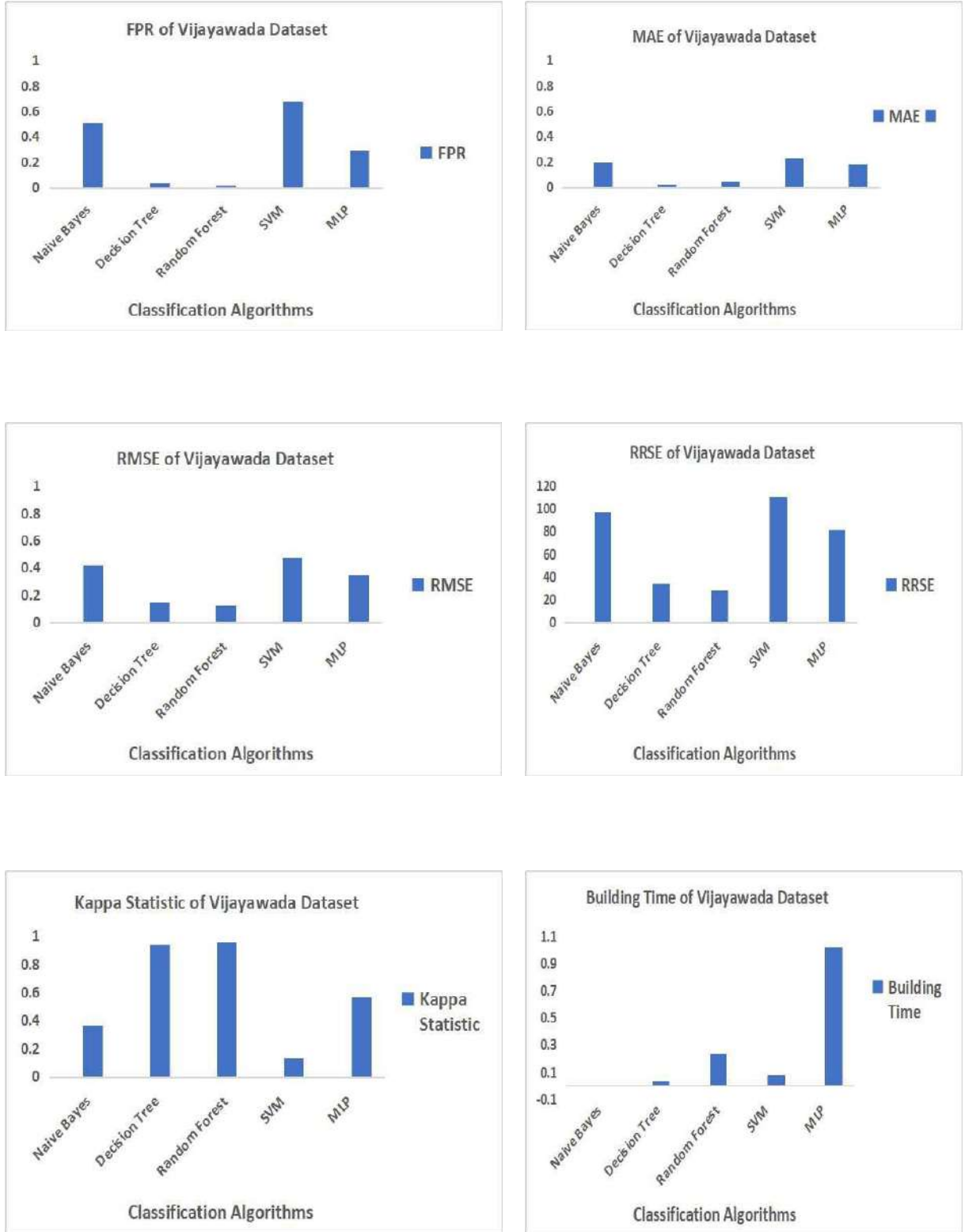


Fig. 10. Error comparison of Vijayawada dataset

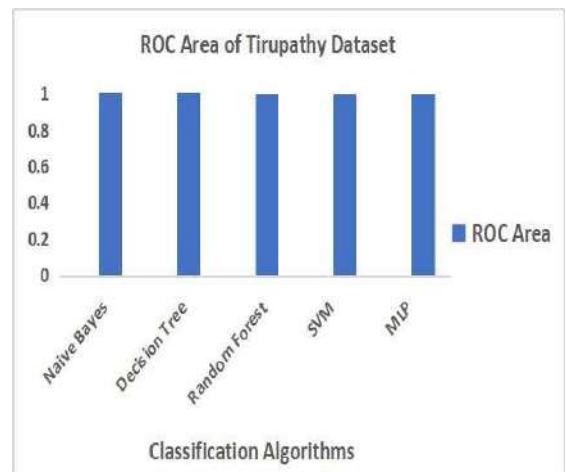
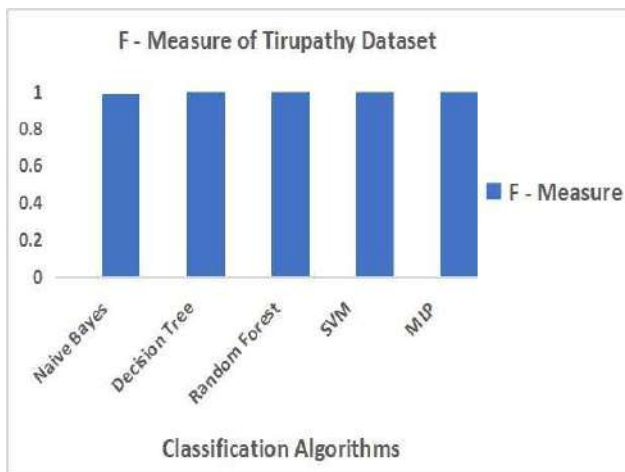
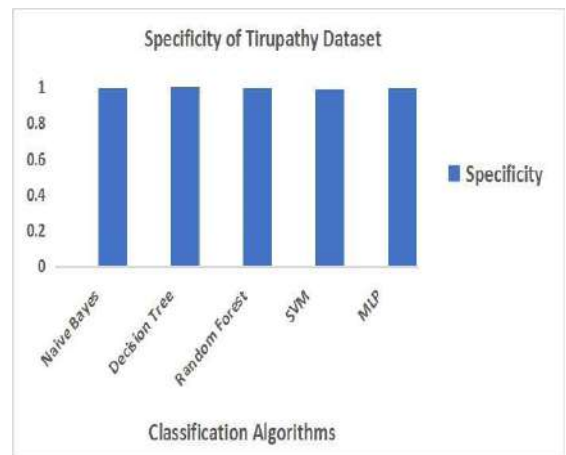
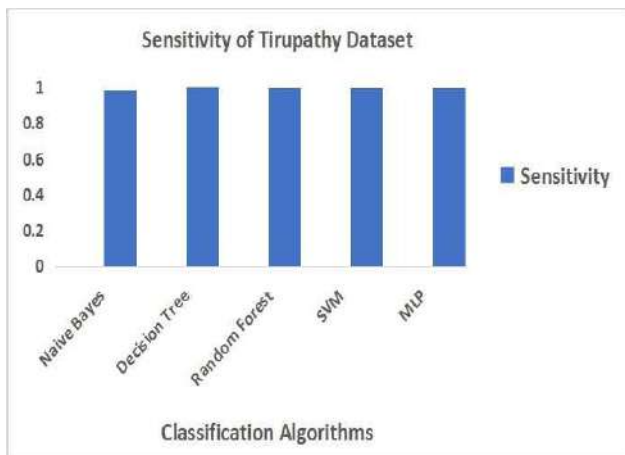
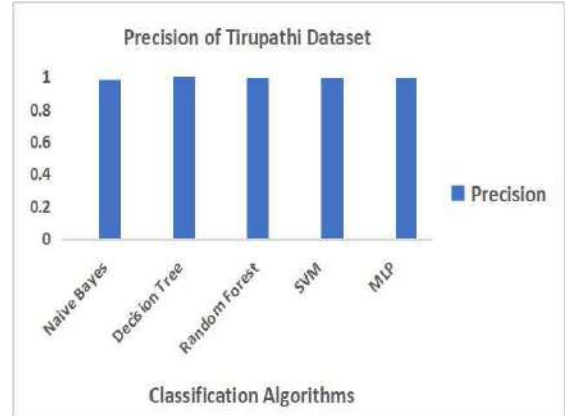
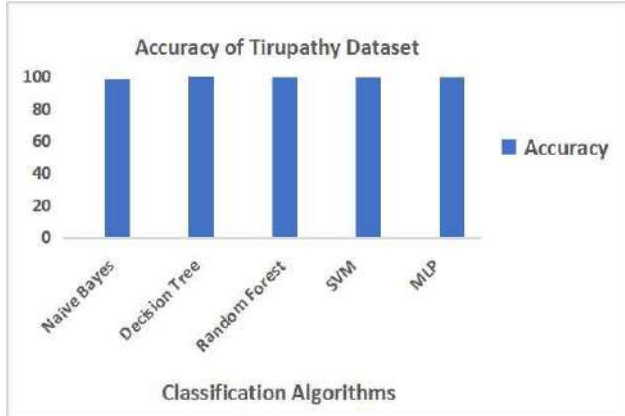


Fig. 11. Performance comparison of Tirupathi dataset

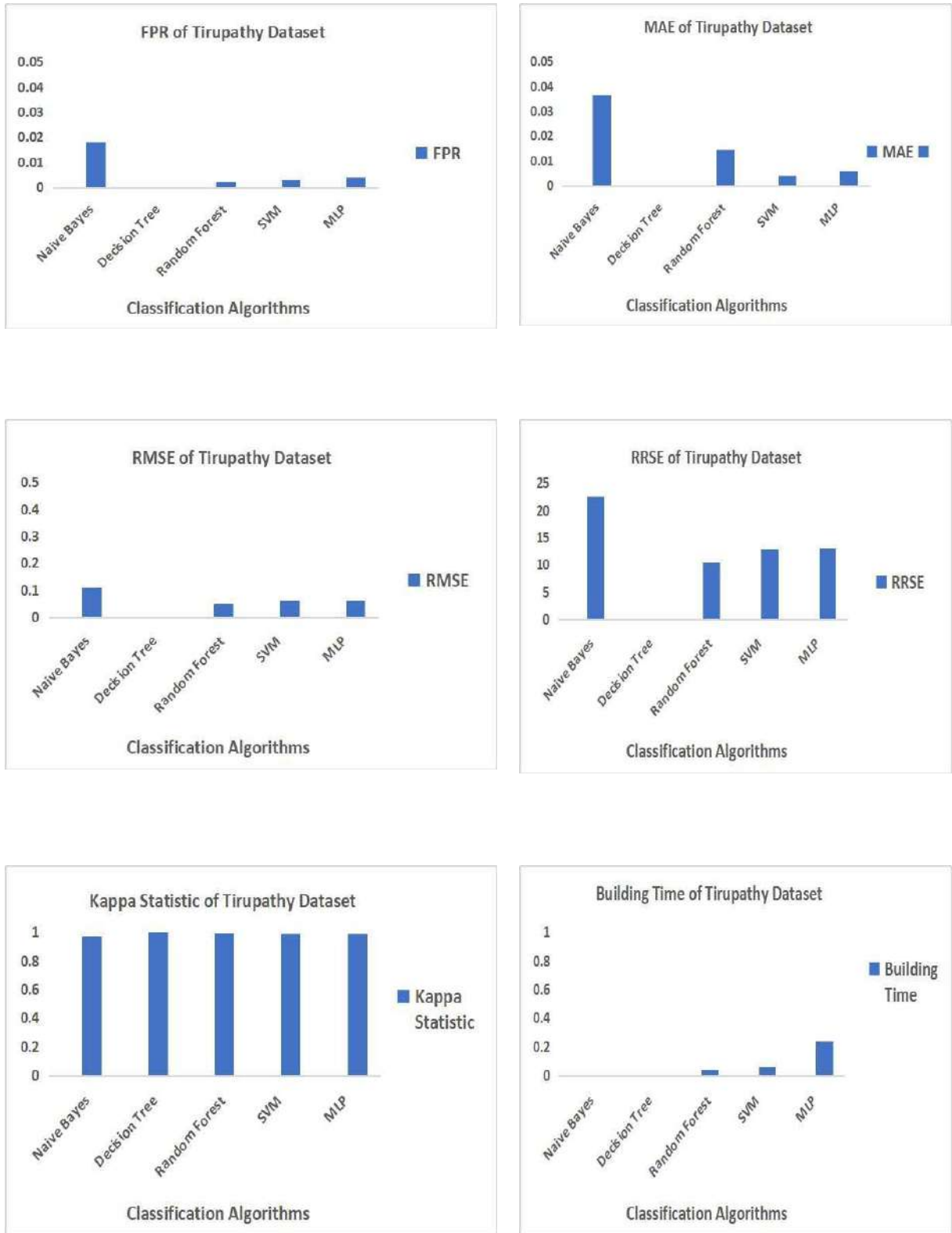


Fig. 12. Error comparison of Tirupathi dataset

Accuracy, Precision, Sensitivity and Specificity are very high in Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets and subsequently error measures are very less for the same datasets. Accuracy, Precision, Sensitivity and Specificity are very high in Random Forest classification algorithm for Vijayawada dataset. Building time is more for MLP than other classifiers in all the three datasets. Building time for MLP in Vijayawada dataset is more than Visakhapatnam and Tirupathi dataset. This may be due to more no of records in Vijayawada dataset than other datasets.

3.6: Conclusions

In this experimentation, Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron Classification Algorithms were considered for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time in classifying liver patients dataset.

Classification performance is very high in Decision Tree classification algorithm for Visakhapatnam and Tirupathi datasets, where as Classification performance is very high in Random Forest classification algorithm for Vijayawada dataset. Building time is more for MLP in Vijayawada dataset.

3.7: Future Scope

The performance of classification algorithms may be improved by selecting important features in classification of liver disease diagnosis. It can also be enhanced by ensembling the classifiers.

CHAPTER 4

Analysis of Geographical effect of various regions on Liver disease

4.1: Introduction

Statistical Analysis plays a significant role in population comparison to investigate the geographical effect on liver diseases. In this study the common attributes were considered from the three datasets for the population comparison. Three data sets were evaluated using analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). ANOVA statistical analysis applied on each single dependent variable that are ALP, DB, SGOT, SGPT and TB to test the significant difference among three datasets. MANOVA statistical analysis applied on all combinations of common attributes to envisage the significance among three datasets.

4.2: Statistical Analysis

Statistical Analysis involves the study of methods for collecting, summarizing, and interpreting data. Statistics formalizes the process of making decisions. The applications of statistics in sciences, economics, computer science, finance, psychology, sociology, criminology, and many other fields. It will examine a number of ways to investigate the relationships between various characteristics of data. Statistics represent, how to organize and display data, and how to test the data to make effective conclusions.

In this study Standard statistical methods ANOVA and MANOVA are applied to evaluate the significance between two populations for better classification. ANOVA is used to test the significant difference in a single dependent variable among two or more groups formed by a single independent or classification variable, whereas MANOVA is used to test the significant difference in more than one dependent variable and several independent variables.

Three Liver patient datasets were used in this study, one is collected from North Coastal Andhra Pradesh i.e Visakhapatnam, Central

Andhra Pradesh i.e Vijayawada and Rayalaseema i.e Tirupathi. The attributes in these datasets are represented in the table 6.

Table. 6. List of attributes of regional datasets

Regional Dataset	Visakhapatnam	Vijayawada	Tirupathi
Attribute No			
1	AGE	AGE	AGE
2	GENDER	GENDER	GENDER
3	TB	TB	TB
4	DB	DB	DB
5	SGOT	AST (SGOT)	SGOT
6	SGPT	ALT (SGPT)	SGPT
7	ALP	ALP	ALP
8	-	IB	IB
9	-	SP (TP)	TP
10	-	SA (Albumin)	Albumin
11	-	SG (Globulins)	Globulins
12	-	A/G RATIO	A/G RATIO

The common attributes from the three data sets AGE, GENDER, TB, DB, SGOT, SGPT and ALP are considered for the purpose of population comparison to analyze the geographical effect. In this Group 1 indicates Visakhapatnam dataset, Group 2 indicates Vijayawada dataset and Group 3 indicates Tirupathi dataset.

4.2.1: One way analysis of variance (ANOVA):

ANOVA is used to test the significant difference in a single dependent variable among two or more groups. The F statistics obtained from ANOVA only tell us whether there is any significant difference in the mean values of the three groups.

Between Groups:- Between groups indicates the variability due to the place of data (between groups variability).

$$(\bar{x}_i - \bar{x})^2$$

Within Groups:-With in groups indicates variability due to random error.

$$(x_{ij} - \bar{x}_i)^2$$

Total:- Indicates total variability

The ANOVA F-statistic is a ratio of the Between Group Variation divided by the Within Group Variation.

4.2.2: Multivariate Analysis of variance (MANOVA):

A MANOVA is used to test the hypothesis that one or more independent variables (IVs) or factors, have an effect on a set of two or more dependent variables (DVs). The goal of our analysis is to look for an effect of one or more IVs on several DVs at the same time. Four different multivariate tests were considered to identify the significant effect of the IVs on all of the DVs, as a group.

Descriptive Statistics table provides the minimum, maximum, mean and standard deviation for the independent variables. Descriptive statistics are categorized into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables, and the kurtosis and skewness.

4.3: Results and Discussion

In this ANOVA analysis ALP, DB, SGOT, SGPT and TB were considered as dependent variables and Group was considered as factoring variable. The results of ANOVA were represented in three rows.

The output of ANOVA analysis on ALP, DB, SGOT, SGPT and TB was represented in table 7, table 8, table 9, table 10 and table 11 respectively which indicates whether there is a statistically significant difference between group means.

Table. 7: ANOVA on ALP between three datasets

ANOVA - ALP	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	382.012	226	1.690	6.057	.000
Within Groups	311.154	1115	.279		
Total	693.165	1341			

Table. 8: ANOVA on DB between three datasets

ANOVA - DB	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	279.482	46	6.076	19.019	.000
Within Groups	413.684	1295	.319		
Total	693.165	1341			

Table. 9: ANOVA on SGOT between three datasets

ANOVA - SGOT	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	257.506	127	2.028	5.650	.000
Within Groups	435.659	1214	.359		
Total	693.165	1341			

Table. 10: ANOVA on SGPT between three datasets

ANOVA - SGPT	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	179.806	129	1.394	3.288	.000
Within Groups	513.323	1211	.424		
Total	693.129	1340			

Table. 11: ANOVA on TB between three datasets

ANOVA - SGPT	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	141.279	61	2.316	5.372	.000
Within Groups	551.886	1280	.431		
Total	693.165	1341			

Significance value (P-value) Indicates the probability of getting a mean difference between the groups as high as what is observed by chance. The lower the p-value, the more significant the difference between the groups. The p-value in all the tables of ANOVA analysis is less than 0.05 ($p < 0.05$) can safely reject the null hypothesis that indicates there is more significant difference between groups. Then the populations differ alot on ALP, DB, SGOT, SGPT and TB. So far, ANOVA explores statistically significant differences between the groups as a whole. Then there is need of MANOVA analysis for Multiple Comparisons, to explore which groups differed from each other.

Descriptive Statistics of Visakhapatnam dataset, descriptive Statistics of Vijayawada dataset and descriptive Statistics of Tirupathi dataset are represented in table 12, table 13 and table 14 respectively.

Table 12. Descriptive Statistics of Visakhapatnam Dataset

Attribute	Minimum	Maximum	Mean	Standard Deviation
AGE	5	68	29.207	12.226
TB	0.3	18.2	1.474	2.172
DB	0.1	13	0.604	1.514
SGOT	14	175	40.626	26.324
SGPT	10	267	36.247	32.343
ALP	28	605	95.753	87.995

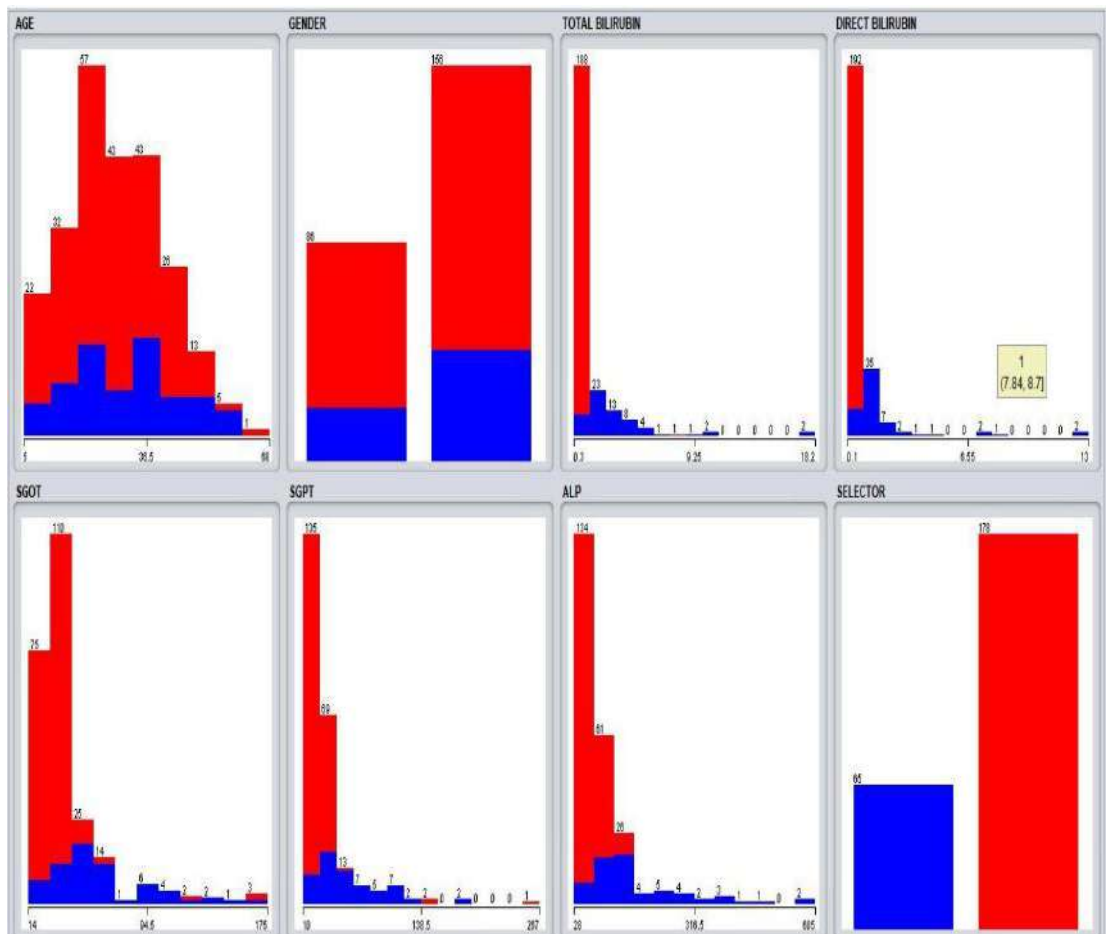


Fig. 13. Descriptive Statistics of Visakhapatnam Dataset

Table 13. Descriptive Statistics of Vijayawada Dataset

AGE	0	90	45.748	17.782
TB	0.1	31.6	0.925	1.886
DB	0.1	22.8	0.447	1.433
SGOT	2	4980	58.523	244.4
SGPT	2	3642	57.552	220.689
ALP	0.7	769	114.74	90.394
IB	0.1	15	0.511	0.979
TP	2.3	502	7.842	20.523
Albumin	0.8	4.9	3.147	.0754
Globulins	12	24	3.735	1.37
A/G RATIO	0.1	2.5	0.903	0.321

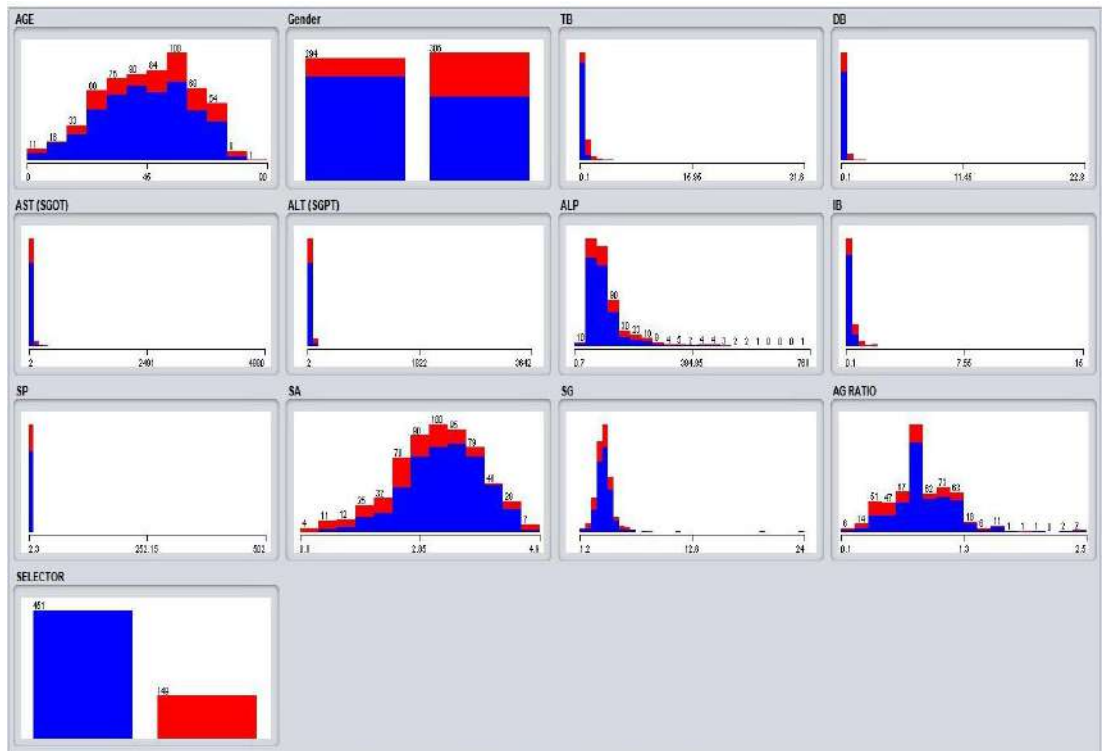


Fig. 14. Descriptive Statistics of Vijayawada Dataset

Table 14. Descriptive Statistics of Tirupathi Dataset

Attribute	Minimum	Maximum	Mean	Standard Deviation
AGE	19	85	47.259	13.905
TB	0.2	1	0.726	0.138
DB	0.1	0.5	0.257	0.054
SGOT	10	45	23.733	7.97
SGPT	11	45	20.657	6.27
ALP	18	188	90.717	19.38
IB	0.1	0.7	0.468	0.128
TP	5.8	72	7.331	2.927
Albumin	2.9	4.8	4.17	0.293
Globulins	2.1	25	3.09	1.053
A/G RATIO	1	2.2	1.365	0.239

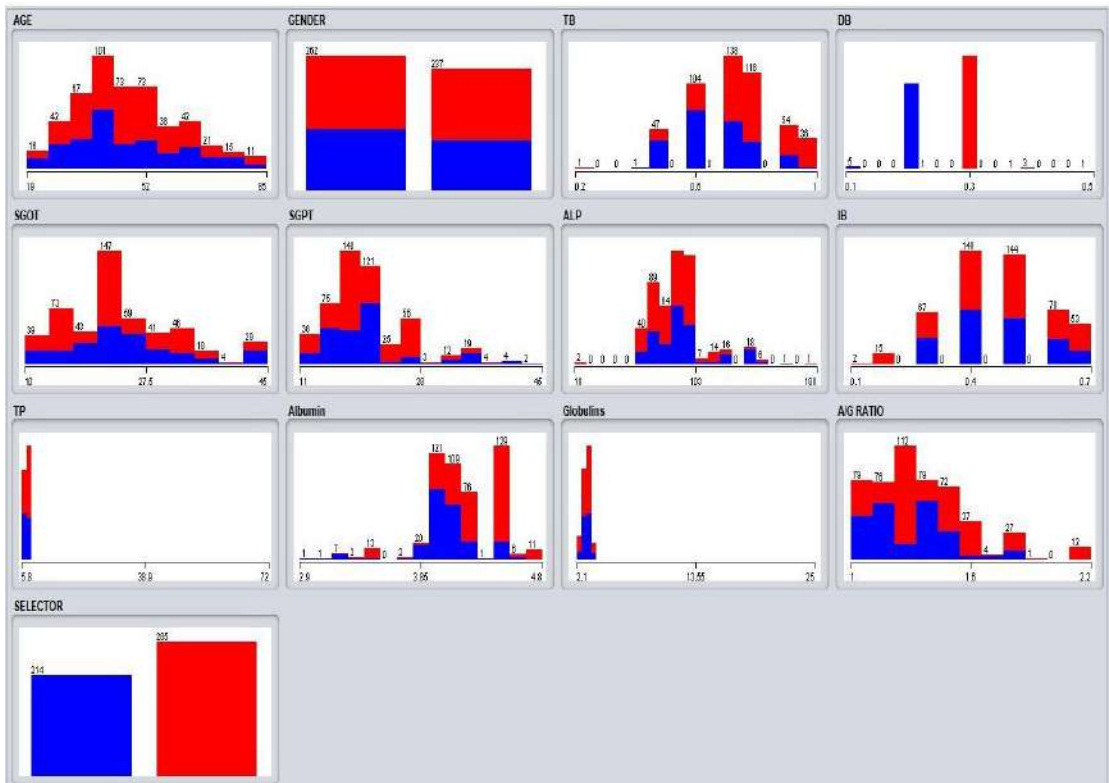


Fig.15. Descriptive Statistics of Tirupathi Dataset

DF: degrees of freedom

Test Statistic: The name of the statistical test shown on this row of the report. The four multivariate tests are followed by the univariate F-tests of each variable.

Test Value: The value of the test statistic.

DF1: The numerator degrees of freedom of the F-ratio corresponding to this test.

DF2: The denominator degrees of freedom of the F-ratio corresponding to this test.

F-Ratio: The value of the F-test corresponding to this test. In some cases, this is an exact test. In other cases, this is an approximation to the exact test.

Prob Level: The significance level of the above F-ratio. The probability of an F-ratio larger than that obtained by this analysis. For example, to test at an alpha of 0.05, this probability would have to be less than 0.05 to make the F-ratio significant.

Decision (0.05): The decision to accept or reject the null hypothesis at the given level of significance. Note that you specify the level of significance when you select Alpha.

Unlike the univariate situation in which there is only one statistical test available (the F-ratio), the multivariate situation provides several alternative statistical tests. These tests are Wilks' Lambda, Lawley's trace, Pillai's trace and Roy's largest root. When the hypothesis degrees of freedom, h , is one, all four test statistics will lead to identical results. When $h > 1$, the four statistics will usually lead to the same result.

Wilks' Lambda, Lawley's trace, and Roy's largest root are often more powerful than Pillai's trace if $h > 1$ and one dimension accounts for

most of the separation among groups. Pillai's trace is more robust to medical analysis from assumptions than the other three.

The following assumptions are made when using a MANOVA.

1. The response variables are continuous.
2. The residuals follow the multivariate-normal probability distribution with means equal to zero.
3. The variance-covariance matrices of each group of residuals are equal.
4. The individuals are independent.

Multivariate analysis of variance (MANOVA) is an extension of common analysis of variance (ANOVA). In ANOVA, differences among various group means on a single-response variable are studied. In MANOVA, the number of response variables is increased to two or more. The hypothesis concerns a comparison of vectors of group means.

The general purpose of multivariate analysis of variance (MANOVA) is to determine whether multiple levels of independent variables on their own or in combination with one another have an effect on the dependent variables.

MANOVA examines the degree of variance within the independent variables and determines whether it is smaller than the degree of variance between the independent variables. If the within subjects variance is smaller than the between subjects variance it means the independent variable has had a significant effect on the dependent

In this study the common attributes from the three datasets TB, DB, SGOT, SGPT, ALP are considered for the multivariate analysis. Multivariate tests and their significant values (p) for the combination of attributes at different levels are represented in Table 15.

Table 15. Multivariate Tests Significance (at p<.005 level)

Level	variables	Pillai's Trace Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
2-way Interactions	DB*ALP	.629	1133.53	2.000	1338.000	.000	.629
	DB*SGOT	.135	104.031	2.000	1338.000	.000	.135
	DB*SGPT	.137	106.212	2.000	1337.000	.000	.137
	SGOT*ALP	.628	1127.60	2.000	1338.000	.000	.628
	SGPT*ALP	.628	1126.30	2.000	1337.000	.000	.628
	SGOT*SGPT	.065	46.306	2.000	1337.000	.000	.065
	TB*ALP	.630	1138.87	2.000	1338.000	.000	.630
	TB*DB	.399	443.812	2.000	1338.000	.000	.399
	TB*SGOT	.286	268.204	2.000	1338.000	.000	.286
	TB*SGPT	.288	270.857	2.000	1337.000	.000	.288
3-way Interactions	DB*SGOT*ALP	.631	761.415	3.000	1337.000	.000	.631
	DB*SGOT*SGPT	.142	73.857	3.000	1336.000	.000	.142
	SGOT*SGPT*ALP	.628	751.625	3.000	1336.000	.000	.628
	TB*DB*SGOT	.407	306.048	3.000	1337.000	.000	.407
	TB*DB*SGPT	.410	308.974	3.000	1336.000	.000	.410
4-way Interactions	DB*SGOT*SGPT*ALP	.631	571.128	4.000	1335.000	.000	.631
	TB*DB*SGOT*ALP	.682	717.477	4.000	1336.000	.000	.682
	TB*SGOT*DB*SGPT	.392	215.767	4.000	1337.000	.000	.392
5-way Interaction	TB*DB*SGOT*SGPT*ALP	.683	574.015	5.000	1334.000	.000	.683

The combination of common attributes at different levels are 2-way Interactions, 3-way Interactions, 4-way Interactions and 5-way Interactions. The combination of attributes at 2-way Interactions are DB*ALP, DB*SGOT, DB*SGPT, SGOT*ALP, SGPT*ALP, SGOT*SGPT, TB*ALP, TB*DB, TB*SGOT and TB*SGPT.

The combination of attributes at 3-way Interactions are DB*SGOT*ALP, DB*SGOT*SGPT, SGOT*SGPT*ALP, TB*DB*SGOT and TB*DB*SGPT.

The combination of attributes at 4-way Interactions are DB*SGOT*SGPT*ALP, TB*DB*SGOT*ALP and TB*SGOT*DB*SGPT.

The combination of attributes at 5-way Interaction is TB*DB*SGOT*SGPT*ALP.

The significance level defines the distance the sample mean must be from the null hypothesis to be considered statistically significant. The confidence level defines the distance for how close the confidence limits are to sample mean. In this study, significance level is 0.05, the corresponding confidence level is 95%. If the P value is less than your significance (alpha) level, the hypothesis test is statistically significant. If the confidence interval does not contain the null hypothesis value, the results are statistically significant.

The Significant values for all the combinations of common attributes in multivariate analysis depicted in table 15 are 0.000 which is less than 0.05 ($p < 0.05$) can safely reject the null hypothesis that indicates there is more significant difference between groups. Then we can say that populations differ a lot on all the combinations that are DB*ALP, DB*SGOT, DB*SGPT, SGOT*ALP, SGPT*ALP, SGOT*SGPT, TB*ALP, TB*DB, TB*SGOT, TB*SGPT, DB*SGOT*ALP, DB*SGOT*SGPT, SGOT*SGPT*ALP, TB*DB*SGOT, TB*DB*SGPT,

DB*SGOT*SGPT*ALP, TB*DB*SGOT*ALP, TB*SGOT*DB*SGPT and TB*DB*SGOT*SGPT*ALP.

4.4: Conclusions

In this experimentation, the common attributes of the three data sets ALP, DB, SGOT, SGPT and TB are considered for ANOVA and MANOVA. The significance level considered for the analysis is 0.05, the corresponding confidence level is 95%. The Significant values in the ANOVA analysis for all the common attributes are less than 0.05 ($p < 0.05$) can safely reject the null hypothesis that indicates there is more significant difference between groups. Then the populations differ alot on ALP, DB, SGOT, SGPT and TB. The Significant values in the MANOVA analysis for all the combinations of common attributes are 0.000 which is less than 0.05 ($p < 0.05$) can safely reject the null hypothesis that indicates there is more significant difference between groups. The results indicates that there is more significant difference among three liver datasets that means there is a geographical effect on liver disesses. This may be due to food habits, alcoholic consumption, air pollution, life style etc. Then there is a need of localized modifications for the identification of liver diseases.

4.5: Future Scope

ANOVA and MANOVA analysis is also suggested for the various confidence levels like 99 % and 90 %. This statistical analysis may be applied for various regions of India i.e different states of India to investigate the geographical effect and to suggest the localized settings for the diagnosis of liver diseases.

CHAPTER 5

CONCLUSIONS

5.1: Conclusions:

Liver disease is the tenth most common cause of death in India. An early liver disease diagnosis will decrease the mortality rate. Naive Bayes, Decision Tree, Random Forest, Support Vector Machines and Multi Layer Perceptron Classification Algorithms were considered for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity, Specificity, F-Measure, ROC-Area, FPR, MAE, RMSE, RRSE, Kappa Statistic and Building Time on the three datasets in the liver disease diagnosis. Liver disease diagnosis indicates the difference in classification performance of various classifiers with different region data sets.

Statistical Analysis was proposed to analyze the liver patient's populations of various regions of Andhra Pradesh to find the reason for the difference.

ANOVA and MANOVA statistical analysis methods are applied on the common attributes of these three datasets and results will indicate that there is more significant difference among three liver datasets that means there is a geographical effect on liver diseases. Then there is a need of localized modifications for the identification of liver diseases. Mobile Apps plays a major role for the early diagnosis of Liver disease.

5.2: Future Scope:

The performance of classification algorithms may be improved by using feature selection for selecting important features in the classification of liver disease diagnosis. It can also be enhanced by ensembling of classifiers.

ANOVA and MANOVA analysis is also suggested for the various confidence levels like 99 % and 90 %. This statistical analysis may be applied for various regions of India i.e different states of India to investigate the geographical effect and to suggest the localized settings for the diagnosis of liver diseases.

References

1. Lung-Cheng Huang, Sen-Yen Hsu and Eugene Lin: “A Comparison Of Classification Methods For Predicting Chronic Fatigue Syndrome Based On Genetic Data”. In Proceedings of the Journal of Translational Medicine, pages 1-8, 2009.
2. Rafael Berlanga, Ernesto Jimenez-Ruiz, Victoria Nebot and David Manset: “Medical Data Integration and the Semantic Annotation of Medical Protocols”. In Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, pages 644-649, 2008.
3. Kemal Polat , Seral Sahan, Halife Kodaz and Salih Gunes: “Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism”. In Proceedings of the Expert Systems with Applications, pages 172-183, 32, 2007.
4. Humar Kahramanli and Novruz Allahverdi: “Mining Classification Rules for Liver Disorders”. In Proceedings of the International Journal of Mathematics and Computers in Simulation, Vol 3, No1, pages 9-19, 2009.
5. Paul R. Harper: “A Review And Comparison Of Classification Algorithms For Medical Decision Making”. In Proceedings of the Health policy, Elsevier, 71 Conference, pages 315–331, 2005.
6. Rong-Ho Lin: “An intelligent model for liver disease diagnosis”. In Proceedings of the Artificial intelligence in medicine , 47, pages 53-62, 2009.
7. Bendi Venkata Ramana, Prof. M. S. Prasad babu and Prof. N. B. Venkateswarlu: “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”. In the proceedings of

International Journal of Database Management Systems (IJDMS),
Vol.3, No.2, 101-114, May 2011.

8. Rakhshani Fatmehsari and F. Bahrami: "Assessment of Parkinson's Disease: Classification and Complexity Analysis". In Proceedings of the 17th Iranian International Conference on Biomedical Engineering (ICBME '10), pages 3-4, 2010.
9. Mohd Fauzi Othman and Mohd Ariffanan Mohd Basri: "Probabilistic Neural Network For Brain Tumor Classification". In Proceedings of the 2nd IEEE International Conference on Intelligent Systems, Modelling and Simulation, pages 136-138, 2011.
10. Li Bin Ren, Yang Yi and HuiYou Chang: "An Improved Binary Tree SVM Classification Algorithm based on Bayesian". In Proceedings of the IEEE International Asia-Pacific Conference on Information Processing, pages 178-181, 2009.
11. Ying Li and Bo Cheng: "An Improved k-Nearest Neighbor Algorithm and Its Application to High Resolution Remote Sensing". In Proceedings of the 17th IEEE International Conference on Geoinformatics, 2009.
12. Giorgia Macchiavello, Gabriele Moser, Giorgio Boni and Sebastiano B. Serpico: "Automatic Unsupervised Classification Of Snow-Covered Areas By Decision-Tree Classification And Minimum-Error Thresholding". In Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing(IGARSS '09), pages 4842-1000-1003, 2009.
13. Mathur and G. M. Foody: "Multiclass and Binary SVM Classification: Implications for Training and Classification Users". IEEE Geoscience And Remote Sensing Letters, Vol. 5, No. 2, pages 241-245, April 2008.
14. Parviz Zeaiean Firouzabadi: "Performance Evaluation of Supervised Classification of Remotely Sensed Data For Crop

- Acreage Estimation”. In Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS '01), pages 2718-2720, 2001.
15. Yu-guo Wang and Hua-peng Li: “Remote sensing image classification based on artificial neural network: A case study of Honghe Wetlands National Nature Reserve”. In Proceedings of the IEEE International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE'10), pages 17-20, 2010
 16. Zhou Faguo, Yang Bingru, Zhang Fan and Yu Xingang: “Research on Short Text Classification Algorithm Based on Statistics and Rules”. In Proceedings of the 3rd IEEE International Symposium on Electronic Commerce and Security , pages 3-7, 2010.
 17. Zheng-Tao Yu, Lu Han, Cun-Li Mao, Jian-Yi Guo, Xiang-Yan Meng And Zhi-Kun Zhang: “Study on The Construction of Domain Text Classification Model With The Help of Domain Knowledge”. In Proceedings of the 7th IEEE International Conference on Machine Learning and Cybernetics, Kunming, pages 2612-2617, 2008.
 18. Ayse Cufoglu, Mahi Lohi and Kambiz Madani: “A Comparative Study Of Selected Classifiers With Classification Accuracy In User Profiling”. In Proceedings of the IEEE International Conference on Computer Science And Information Engineering, pages 708-712, 2008.
 19. Noor Ezan Abdullah, Hadzli Hashim, Fairul Nazmie Osman and Fardalila Mohd Adam: “Comparison Between Various Supervised ANN Algorithm Using RGB Indices For Plaque Lesion Classification”. In Proceedings of the IEEE International Conference on Electronic Devices, Systems And Applications (ICEDSA2010), pages 236-241, 2010.
 20. B.Surendiran, Y.Sundaraiah, A.Vadivel: “Classifying Digital Mammogram Masses using Univariate ANOVA Discriminant Analysis ”. In Proceedings of the IEEE International Conference

- on Advances in Recent Technologies in Communication and Computing , pages 175-177, 2009.
21. Mireille Tohmé, Régis Lengellé and Virginie Freytag: “A multiclass multivariate Mireille group comparison test: Application to drug safety”. In Proceedings of the 32nd IEEE International Conference on EMBS, 2006, pages 4711-4714, September 2010.
 22. Bendi Venkata Ramana, Prof. M. S. Prasad babu and Prof. N. B. Venkateswarlu: “A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis”, International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 3, No 2, ISSN: 1694-0814, PP 506-516, May 2012.
 23. Z. A. Dastgheib, B. Lithgowand Z. Moussavi: “Application of Fractal Dimension on Vestibular Response Signals for Diagnosis of Parkinson’s Disease”. In Proceedings of the 33rd IEEE International Conference on EMBS, pages 7892-7895, September 2011.
 24. S. Dimitrova: “Investigations of Some Human Physiological Parameters in Relation to Geomagnetic Variations of Solar Origin and Meteorological Factors ”. In Proceedings of the 2nd IEEE International Conference on Recent Advances in Space Technologies, pages 728-733, 2005.
 25. <http://www.thehindu.com/todays-paper/tp-national/tp-andhrapradesh/watch-your-weight-keep-fatty-liver-disease-at-bay/article2824434.ece>
 26. <http://www.newindianexpress.com/lifestyle/health/India-May-Become-World-Capital-of-Liver-Diseases/2014/04/19/article/2177734.ece>

Appendix

Liver Diagnosis App

Liver Diagnosis App was developed by using Decision Tree classifier for the benefit of the liver patient to know the liver functionality and that is available in the Play store. By using this App first they have to observe the symptoms represented in Fig. 16 that are Loss of Weight, Abdomen Pain, Mass for per Abdomen, Loss of Appetite, Tiredness, Severe Itching, Fatigue, Weakness and Dark Urine .



Fig. 16. Liver Disease Symptoms

If they have most of the symptoms then and proceed for the Liver Function Tests (LFTs). LFTs are simple blood tests through which it can easily diagnose status of the liver. Then they have to enter their LFT values with Age and Gender through this App as shown in Fig. 17 and finally click the diagnosis for the diagnosis of the liver disease.

The screenshot shows a mobile application interface for "Liver Diagnostics App". The title bar is blue with the text "Liver Diagnostics App" in white. Below the title bar, the text "LIVER FUNCTION TEST" is displayed in bold blue letters. The form contains several input fields and a dropdown menu. At the top, there is a field for "Enter Age" with a red underline. Below it is a dropdown menu currently showing "Male". The form lists seven liver function tests, each with a red label, an input field, and a reference range in brackets:

Test	Input Field	Reference Range
TB	Enter TB	[0.22 - 1.0]
DB	Enter DB	[0.0 - 0.2]
ALP	Enter ALP	[110 - 310]
SGPT	Enter SGPT	[Up to 45]
SGOT	Enter SGOT	[5 - 40]
TP	Enter TP	[5.5 - 8]
ALB	Enter ALB	[3.5 - 5]

The bottom of the screen shows the standard Android navigation bar with back, home, and recent apps buttons.

Fig. 17. Liver Function Tests

Finally the disease conformation is shown in Fig. 18. The liver may be either Direct Bilirubinemia (or) Indirect Bilirubinemia (or) Hepatocellular Pattern (or) Cholestatic Pattern. The patient suffering with any of this disease will be treated as liver patient and classified as Liver patient is Yes.

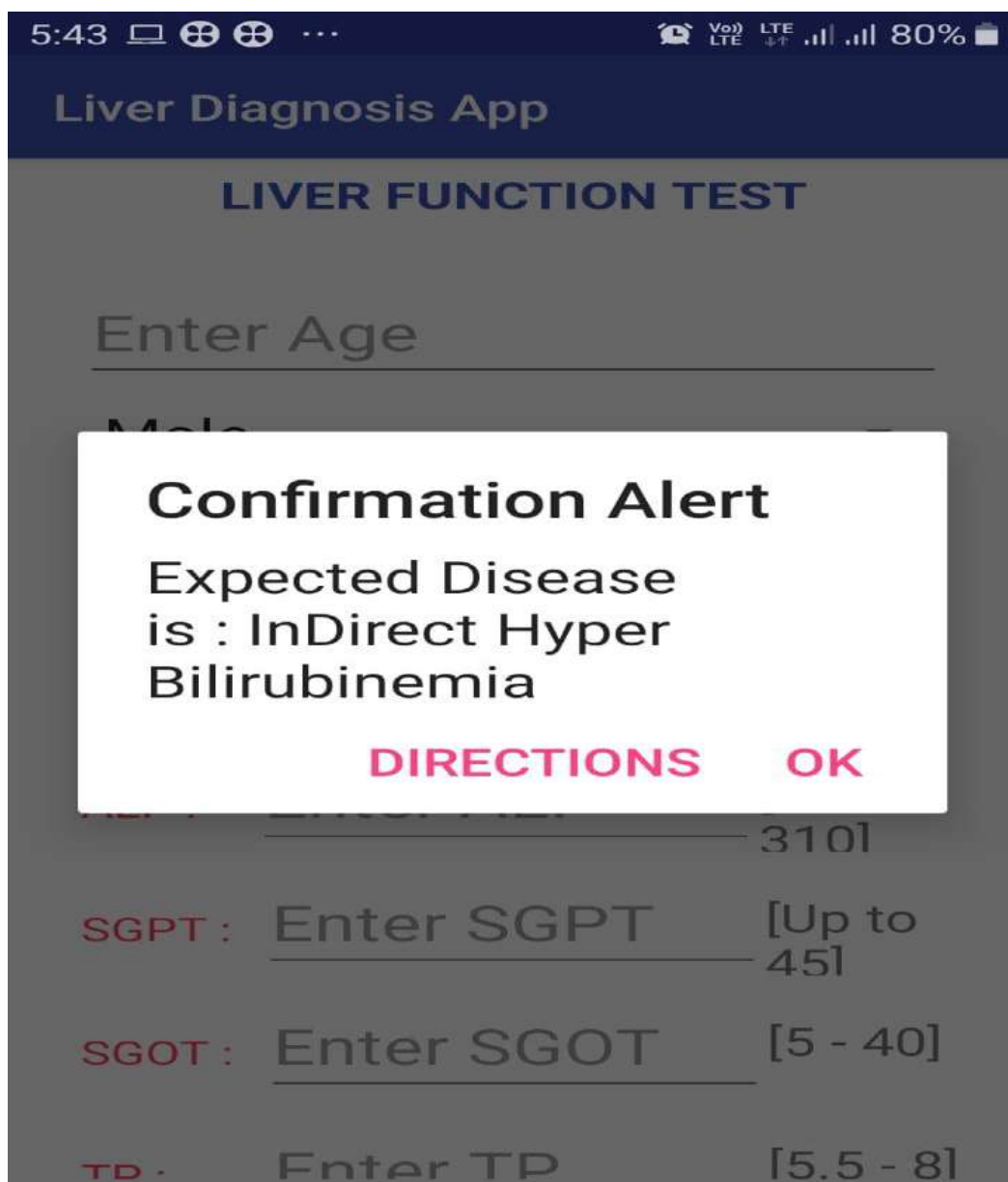


Fig. 18. Liver Disease Diagnosis Conformation

Liver Diagnosis App gives a very good support for patients to diagnose liver disease on their own and also it will reduce the unnecessary waiting at the hospital. After the conformation of the disease patient can meet the doctor for the necessary medication. Liver Diagnosis App can also helps the doctors to reduce the crowd at the hospital.