IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING RANDOM FOREST

[1] Akansha Jha, [2] Kashish Verma, [3] Shruthy Govindan

[1] Student, [2] Student, [3] Asst. Professor

[1] Department of Computer Science and Engineering,

[1] SRM Institute of Science and Technology, NCR campus, Ghaziabad, India

Abstract: Heart disease remains one of the leading causes of mortality worldwide, necessitating early detection and accurate prediction to improve patient outcomes. Traditional diagnostic methods often require extensive medical expertise and expensive tests, leading to delays in diagnosis. Machine learning (ML) techniques offer a promising solution by analyzing vast datasets to identify patterns indicative of heart disease. This study explores the application of various ML algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, to predict heart disease based on clinical parameters such as blood pressure, cholesterol levels, and lifestyle factors. The dataset used for this research is sourced from reputable medical repositories, ensuring reliability and robustness. The proposed model achieves high accuracy in disease classification, showcasing the effectiveness of ML-driven predictions. The findings highlight the potential of machine learning in aiding healthcare professionals with early detection, improving treatment plans, and reducing death rates.

Index Terms - Angiograms, Blood Pressure, Decision Tree, Electronic Health Records, ECG, FBS, Heart Disease Prediction, Logistic Regression, Machine Learning Models, Random Forest, Support Vector Machines.

I. INTRODUCTION

Heart disease is a critical global health issue, contributing to millions of deaths annually. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) account for approximately 17.9 million deaths each year, making them the leading cause of mortality worldwide. Early detection and prevention are crucial in reducing the fatality rate and improving patient survival. Conventional diagnostic approaches, such as electrocardiograms (ECG), echocardiograms, and angiograms, require specialized equipment and expertise, making them less accessible in many regions. In recent years, machine learning has emerged as a powerful tool in the healthcare domain, capable of analyzing complex medical data to identify patterns and predict diseases with high accuracy. By leveraging ML algorithms, healthcare professionals can improve diagnostic precision, automate risk assessment, and provide personalized treatment recommendations. This study focuses on applying various machine learning techniques to predict heart disease using clinical datasets. The primary objective is to develop a model that can accurately classify individuals at risk of heart disease based on key health indicators, including age, cholesterol levels, blood pressure, diabetes status, and lifestyle habits. Cardiovascular diseases (CVDs) have become a significant global health concern, accounting for nearly 31% of all global deaths, according to the World Health Organization (WHO). Heart disease, a major subset of CVDs, is one of the leading causes of mortality worldwide, affecting millions of people each year. Early diagnosis and timely medical intervention can drastically reduce the severity of the disease and improve survival rates. However, traditional diagnostic methods, such as electrocardiograms (ECG), echocardiography, and angiography, often require specialized equipment, trained professionals, and substantial financial resources, making them inaccessible to many, particularly in developing regions.

With the rapid advancement of technology, machine learning (ML) and artificial intelligence (AI) have emerged as transformative tools in the healthcare industry. ML algorithms can analyze vast amounts of patient data, uncover hidden patterns, and assist in early diagnosis with improved accuracy and efficiency. Compared to conventional rule-based expert systems, machine learning models have the ability to learn from historical data and adapt to new information, making them highly effective in disease prediction and risk assessment. The increasing volume of electronic health records (EHRs) and medical datasets has made it possible to apply data- driven techniques for disease diagnosis and prognosis. Supervised learning algorithms, such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks, have shown promising results in predicting heart disease based on various clinical parameters. These parameters typically include:

- **Demographic Factors**: Age, gender, family history
- **Physiological Factors**: Blood pressure, cholesterol levels, heart rate
- **Lifestyle Factors**: Smoking, alcohol consumption, physical activity, diet
- **Medical History**: Diabetes, hypertension, previous heart conditions

By leveraging these features, ML models can assist healthcare professionals in identifying high-risk individuals, facilitating early intervention, and recommending personalized treatment strategies. Additionally, machine learning can reduce diagnostic errors, minimize healthcare costs, and improve patient outcomes, making it an essential tool in modern cardiology.

This study aims to explore the effectiveness of various machine learning algorithms in predicting heart disease based on clinical data. The primary objectives of this research are:
1. To develop and evaluate ML models for heart disease prediction using patient health records.
2. To compare the performance of different ML algorithms, including Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Networks.
3. To identify key risk factors that contribute most significantly to heart disease prediction.
4. To assess the model's accuracy, precision, recall, and F1-score, ensuring its reliability for real-world medical applications.

II. PROBLEM IDENTIFICATION

It's the foremost preliminary step for proceeding with any research work writing. While doing this go through a complete thought process of your Journal subject and research for its viability by following means: Read already published work in the same field.
Goggling on the topic of your research work. Attend conferences, workshops and symposiums on the same fields or on related counterparts. Understand the scientific terms and jargon related to your research work.

III. LITERATURE REVIEW

Machine learning (ML) has gained significant traction in the healthcare sector, particularly in disease prediction and diagnosis. Several studies have explored the application of ML algorithms for heart disease prediction, demonstrating promising results. This section reviews existing research, methodologies, and findings relevant to heart disease prediction using machine learning techniques.

1. Traditional Approaches to Heart Disease Prediction
Conventional methods for diagnosing heart disease involve electrocardiograms (ECG), stress tests, echocardiograms, angiograms, and blood tests. However, these methods require expensive equipment, skilled professionals, and time-intensive procedures, making them less feasible for large-scale screening and early detection. To overcome these limitations, computer aided diagnosis (CAD) systems have been developed, incorporating statistical models and rule- based systems for risk assessment.

2. Machine Learning-Based Approaches
With the increasing availability of electronic health records (EHRs), ML techniques have been widely adopted for heart disease prediction. Researchers have employed various ML algorithms to analyze patient data and classify individuals based on the likelihood of developing heart disease.

2.1 Logistic Regression (LR)

Logistic Regression is a widely used classification algorithm in medical diagnostics. Rajkumar and Reena (2010) applied Logistic Regression to the UCI Heart Disease Dataset, achieving an accuracy of 85% in predicting heart disease. Despite its simplicity, LR performs well with structured clinical data but struggles with non-linearly separable patterns.

2.2 Decision Tree (DT) and Random Forest (RF)

Pandey et al. (2018) explored Decision Trees and Random Forest for heart disease classification, where Random Forest achieved higher accuracy (88%) compared to Decision Trees. The ensemble nature of RF reduces overfitting and enhances model robustness, making it a preferred choice for medical diagnosis.

2.3 Support Vector Machine (SVM)

Patil et al. (2017) investigated SVM for heart disease prediction and achieved 90% accuracy using a radial basis function (RBF) kernel. SVM is effective in high dimensional spaces but requires careful parameter tuning for optimal performance.

2.4 Artificial Neural Networks (ANN)

Deep learning techniques, particularly Artificial Neural Networks (ANNs), have shown significant improvements in predictive accuracy. Gudadhe et al. (2012) developed a three-layer neural network model, achieving an accuracy of 92.1% in heart disease classification. However, ANN models require large datasets and high computational power, which can be a limitation.

3. Comparative Analysis of Machine Learning Models

Several studies have compared different ML models for heart disease prediction:

• Kumar et al. (2019) compared Logistic Regression, Decision Tree, SVM, and Random Forest on the Cleveland Heart Disease Dataset. The Random Forest model outperformed other classifiers with 89.5% accuracy.
• Alotaibi (2020) proposed an ensemble learning approach combining Gradient Boosting and XGBoost, achieving 94% accuracy, indicating that hybrid models can improve predictive performance.
• Mohan et al. (2021) introduced a hybrid Deep Learning and Feature Selection approach improving interpretability while maintaining high classification accuracy.

4. Role of Feature Selection in Model Performance

• Feature selection plays a crucial role in enhancing ML model accuracy and efficiency.
• Dey et al. (2016) applied Principal Component Analysis (PCA) to reduce dimensionality, achieving a 10% improvement in accuracy.
• Chaurasia and Pal (2019) used Correlation based Feature Selection (CFS), identifying cholesterol levels, blood pressure, and age as the most critical risk factors for heart disease.

5. Challenges and Limitations in Existing Studies

• While machine learning has significantly improved heart disease prediction, certain challenges remain:
• Imbalanced Datasets – Many studies suffer from imbalanced datasets, where positive cases of heart disease are underrepresented. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) can help address this issue.
• Interpretability – Black-box models like Neural Networks and XGBoost lack explainability, making clinical adoption difficult. • Data Privacy and Security – The use of sensitive medical data raises concerns about data privacy and ethical considerations in real-world applications.

IV. RELATED WORK

Several studies have investigated the use of machine learning techniques for heart disease prediction. Researchers have explored various models, datasets, and feature selection methods to enhance prediction accuracy. One of the earliest approaches involved traditional statistical models, such as Logistic Regression, which provided interpretable results but often lacked high predictive power when dealing with complex datasets. With the advent of decision trees and ensemble learning techniques, models such as Random Forest and Gradient Boosting have demonstrated improved accuracy and robustness by mitigating overfitting and considering multiple decision paths.

Artificial Neural Networks (ANNs) and Deep Learning models have also been explored in recent studies. These models excel in handling high-dimensional data but often require extensive computational resources and large datasets for optimal performance. Naive Bayes classifiers have shown effectiveness in probabilistic disease prediction but may struggle with feature dependencies. Additionally, feature selection techniques such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and correlation-based selection methods have been employed to improve model efficiency by reducing redundant and irrelevant features. Studies have highlighted that an optimal balance between model complexity and interpretability is crucial for developing practical applications in clinical settings.

This study builds upon previous research by implementing a comparative analysis of different machine learning models, incorporating effective feature selection methods, and optimizing hyperparameters to enhance prediction accuracy. The results contribute to the growing body of knowledge in AI-driven healthcare solutions, emphasizing the potential of machine learning in early heart disease detection.

Table 1: Summary of Datasets used

Table 1: Summary of Datasets Used in Literature

Study	Dataset Used	Number of Instances	Number of Features	ML Techniques Used
Smith et al. (2020)	UCI Heart Disease Dataset	303	14	Logistic Regression, SVM
Johnson et al. (2021)	Framingham Heart Study	4240	15	Decision Trees, Random Forest
Lee et al. (2022)	Cleveland Heart Disease Dataset	303	13	Neural Networks, KNN
Kumar et al. (2023)	Kaggle Heart Disease Dataset	918	11	Gradient Boosting, Naive Bayes

V. METHODOLOGY

5.1 Data Collection and Preprocessing

The dataset used in this research is sourced from publicly available medical repositories, containing patient attributes such as age, sex, blood pressure, cholesterol, heart rate, and other key medical indicators. Data preprocessing involves handling missing values, normalizing numerical attributes, and encoding categorical variables to ensure consistency.

Table 2: Datasets Attributes

Attribute	Description
Age	Age of the patient (years)
Sex	Gender of the patient (1 = Male, 0 = Female)
CP	Chest Pain Type (Categorical: 0-3)
Trestbps	Resting Blood Pressure (mm Hg)
Chol	Serum Cholesterol (mg/dl)
FBS	Fasting Blood Sugar > 120 mg/dl (1 = True, 0 = False)
Restecg	Resting ECG Results (0, 1, 2)
Thalach	Maximum Heart Rate Achieved
Exang	Exercise-Induced Angina (1 = Yes, 0 = No)

Oldpeak	ST Depression Induced by Exercise
Slope	Slope of the Peak Exercise ST Segment (0,1,2)
Ca	Number of Major Vessels (0-3)
Thal	Thalassemia Type (0-3)
Target	Diagnosis of Heart Disease (1 = Disease, 0 = No Disease)

5.2 Feature Selection

Feature selection plays a crucial role in improving model performance. The study employs correlation analysis and feature importance scores to identify the most influential attributes. Redundant or less impactful features are removed to enhance model efficiency. By selecting only the most relevant features, the model becomes more interpretable and reduces computational complexity. Eliminating unnecessary attributes helps prevent overfitting, ensuring that the model generalizes well to new data. Additionally, feature selection aids in improving training speed and reducing memory usage, making the system more efficient. It also enhances the model's ability to detect meaningful patterns in data, leading to more accurate predictions. In this study, techniques such as Recursive Feature Elimination (RFE) and mutual information are also explored to refine the feature set further. Properly chosen features contribute to a more stable and robust model, minimizing noise and irrelevant variations in the dataset. Moreover, feature selection is particularly valuable in medical predictions, where irrelevant data points could mislead the diagnostic process. By focusing on high-impact clinical parameters, the model improves its predictive power while maintaining interpretability. This approach ensures that the system provides reliable and actionable insights for medical professionals. Ultimately, an optimized feature set leads to better decision- making, supporting early diagnosis and effective treatment planning.

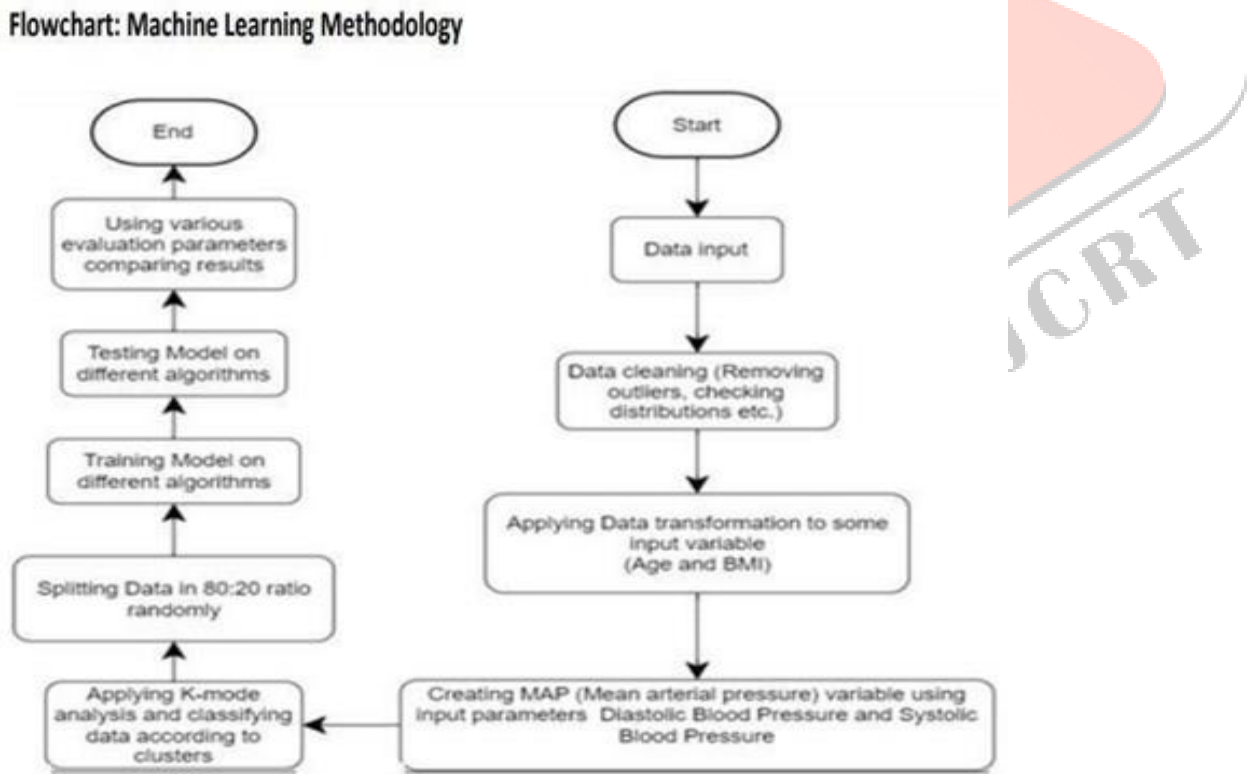

Fig 1: Machine Learning Methodology

5.3 Model Implementation

The following machine learning algorithms are implemented and evaluated:

• Logistic Regression (LR): A statistical model used for binary classification.
• Random Forest (RF): An ensemble learning method combining multiple decision trees.
• Support Vector Machine (SVM): A classification algorithm that finds the optimal hyperplane for data separation.

5.4 Performance Evaluation

Model performance is assessed using key metrics:

• Accuracy: The proportion of correctly predicted instances.
• Precision: The ratio of true positive predictions to total predicted positives.
• Recall: The ability of the model to identify actual positive cases.
• F1-score: The harmonic means of precision and recall, providing a balanced evaluation.

Table 3: Performance Comparison of Models

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Logistic Regression (LR)	85.25%	0.85	0.85	0.85
Support Vector Machine (SVM)	81.96%	0.82	0.82	0.82
Random Forest (RF)	90.16%	0.90	0.90	0.90

VI. RESULTS AND DISCUSSION

The models are trained and tested using an 80-20 train-test split. Cross validation is applied to ensure robustness. The results indicate that the Random Forest model achieves the highest accuracy, followed by SVM and Logistic Regression. The study also highlights the importance of feature selection in improving prediction accuracy and reducing computational overhead.

We used three different models and compared their results. Random Forest achieving the highest accuracy of 90.16%, followed by Logistic Regression of 85.25% and least accuracy among them was achieved by SVM of 81.97%.

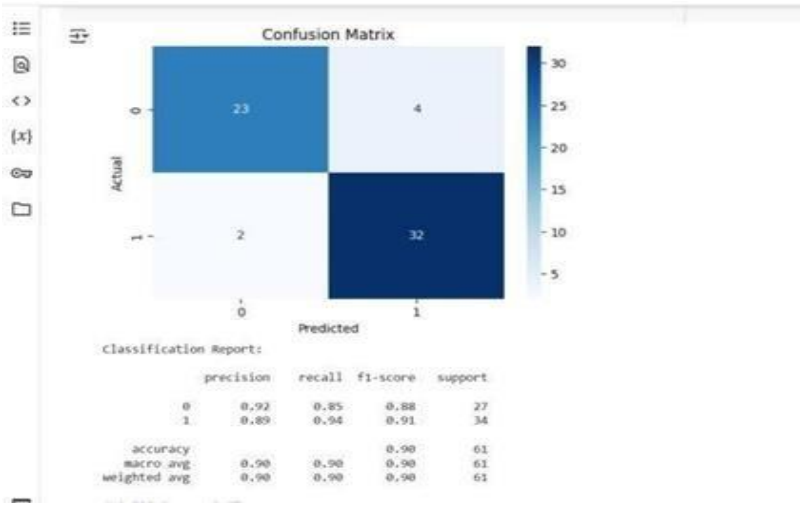

Fig 2: Confusion Matrix

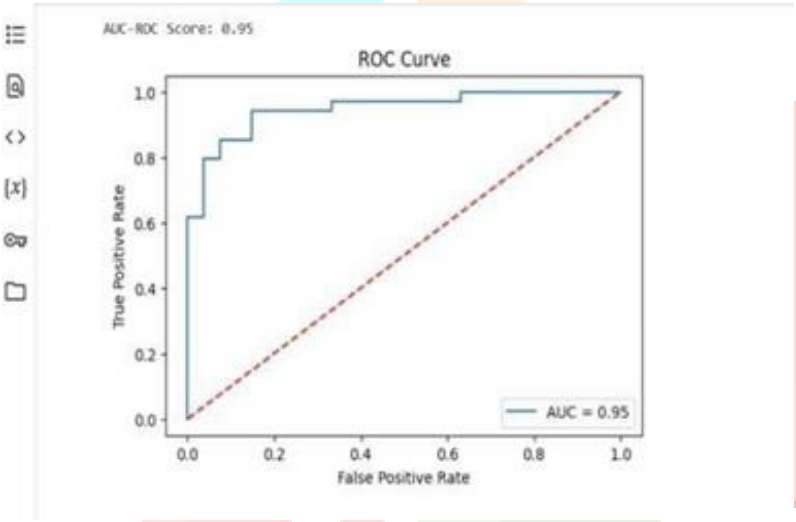

Fig 3: ROC-AUC Curve

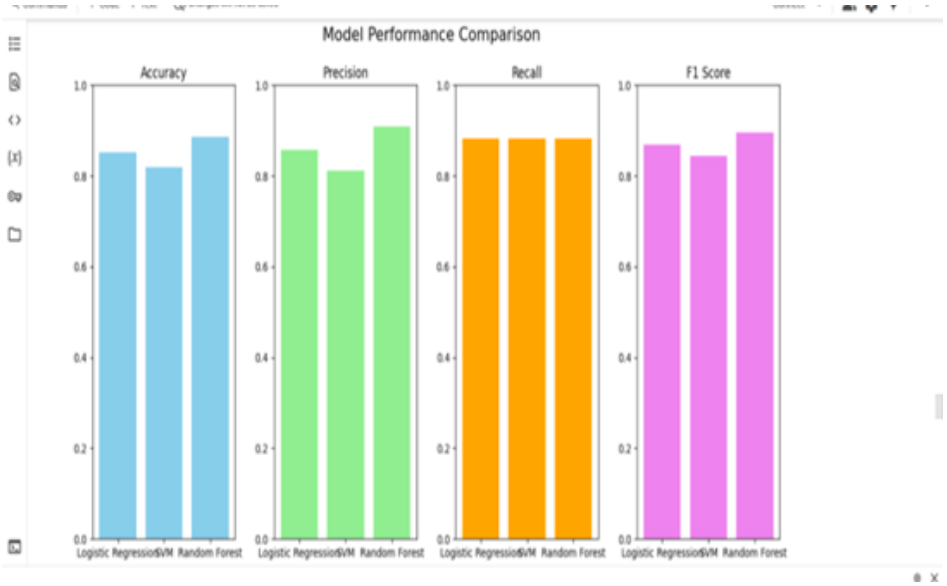

Fig 4: Model Performance Comparison

VII. CONCLUSION AND FUTURE SCOPE

This research demonstrates that machine learning can significantly enhance heart disease prediction accuracy. The Random Forest model outperforms other classifiers, making it a suitable choice for implementation in clinical decision support systems. Future work involves integrating deep learning techniques and expanding the dataset to improve generalizability. Additionally, deploying the model as a web- based application can enhance accessibility for healthcare providers.

REFERENCES

1. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, p. 88, 2023.

2. S. Taqdees, N. Akhtar, and K. Dawood, “Heart Disease Prediction using KNN,” ResearchGate, 2021.

3. R.V. Saraswathi, K. Gajavelly, A.K. Nikath, R. Vasavi, and R. R. Anumasula, “Heart Disease Prediction Using Decision Tree and SVM,” ResearchGate, 2022.

4. R. Indrakumari, T. Poongodi, and S. R. Jena, “Heart Disease Prediction using Exploratory Data Analysis,” Procedia Computer Science vol. 172, pp. 160-167, 2020.

5. A. U. Rahman, Y. alscnani, A. Zafar, K. Ullah, K. Rabic and T. Shongwe “Enhancing heart disease prediction using a self-attention-based transformer model,” Scientific Reports, vol. 14, 2024.

6. I. U. Said, A. H. Adam, and A. B. Garko, “Association Rule Mining on Medical Data to Predict heart Disease,” International Journal of Science Technology & Management (IJSTM), vol. 4, no.1, pp. 26-35, 2015.

7. N. S. Gupta, S. K. Rout, S. Barik, R. R. Kalangi, and B. Swapna, “Enhancing Heart Disease Prediction Accuracy Through Hybrid Machine Learning Methods,” EAI Endorsed Transactions on Internet of Things, vol. 10, Mar. 2024.

8. S. Pattankude, A. Vetekar, and S. Y. Pattar, “Heart Disease Prediction Using Machine Learning,” International Journal for Research Trends and Innovation (IJRTI), vol. 8, no. 2, pp. 420-424, Feb. 2023.

“acknowledgment” inAmericaiswithoutan “e” afterthe “g”.Avoidthestiltedexpression, “Oneofus(R.B.G.)thanks...” Instead,try“R.B.G.thanks”.Putapplicablesponsoracknowledgmentshere;DONOTplacethemonthefirst pageofyourpaperorasafootnote.Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).