PREDICTING THE RISK OF HOSPITLAREADMISSIONSUSING A MACHINE LEARNINGAPPROACH:ACASESTUDYONPATEINTUNDERGO INGSKIN PROCEDURES. A SURVEY

K.PAVANI[1], ,M.DEEPIKA2

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract_ Hospital readmissions continue to be a major concern for healthcare providers across the globe, affecting patient outcomes while placing significant financial and operational pressure on healthcare institutions. In the United States, approximately 20% of Medicare beneficiaries are readmitted to hospitals within 30 days of discharge, contributing to healthcare expenditures of more than $26 billion each year. Despite extensive research on readmission prediction, patients undergoing dermatological and skin-related surgical procedures remain a relatively unexplored group. These patients often face distinct challenges, including post-operative wound care complications, infection risks, and the influence of underlying comorbid conditions.

This research introduces a comprehensive machine learning framework designed to estimate the likelihood of 30-day hospital readmission among patients who have undergone skin-related procedures. The proposed framework encompasses the complete predictive analytics workflow, including data preprocessing, feature extraction, model development, performance evaluation, and interpretability analysis. Four supervised learning algorithms—Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM)—were trained and evaluated using a de-identified electronic health record (EHR) dataset containing both clinical and procedure-specific information.

Experimental findings indicate that the XGBoost model delivers superior predictive performance compared with the other classifiers. The model achieved an AUC-ROC score of 0.921, an accuracy of 91.2%, a precision of 91.5%, and a recall of 88.2%. To improve transparency and support clinical decision-making, SHAP (SHapley Additive exPlanations) was employed to identify the most influential factors contributing to readmission risk. The analysis revealed that previous hospitalization history, the number of existing comorbidities, hospital length of stay, patient age, and the severity of the skin procedure were the most significant predictors.

To facilitate practical clinical adoption, the predictive system has been integrated into an interactive web-based dashboard that allows healthcare professionals to assess patient readmission risk in real time. The platform also supports automated generation of clinical reports, enabling more informed discharge planning and targeted interventions. The proposed framework demonstrates the potential of machine learning and explainable artificial intelligence to enhance readmission prediction, improve patient management strategies, and reduce avoidable healthcare costs..

Index terms - *Hospital Readmission, Machine Learning, XGBoost, Random Forest, Skin Procedures, Electronic Health Records (EHR), SHAP, Risk Prediction, Healthcare Analytics, Feature Engineering*

1. INTRODUCTION

Unplanned hospital readmissions represent one of the most pressing challenges confronting modern healthcare delivery systems. In the United States alone, approximately 3.3 million adult hospital readmissions occur each year, costing an estimated USD 26 billion annually. Similar patterns are observed across Europe, Australia, and rapidly developing healthcare markets such as India, where risingprocedurevolumescoincidewithconstrained post-dischargemonitoring resources.Regulatory bodies such as the Centers for Medicare and Medicaid Services (CMS) have introduced financial penalties for hospitals with excess readmission rates, creating powerful institutional incentives for evidence-based prediction and prevention strategies.

Skin and soft-tissue procedures — encompassing excisions, biopsies, wound debridements, flap repairs, skin grafts, Mohs micrographic surgery, and laser ablations — constitute a significant and growing segment of surgical activity globally. Unlike major organ surgeries, dermatological procedures are frequently performed in ambulatory or day-care settings under the assumption of low post-operative risk. However, emerging evidence indicates that patients undergoing skin procedures carry a non-trivial readmission risk, particularly when comorbidities such as diabetes mellitus, immunosuppression, chronic kidney disease, or peripheral arterial disease are present. Wound complications,includingsurgicalsiteinfections(SSI),dehiscence,andhaematomaformation,account for the majority of preventable readmissions in this cohort.

Traditional risk-stratification tools — notably the LACE Index, the HOSPITAL Score, and the Rothman Index — were developed primarily on general medical or cardiac populations and have demonstrated limited discriminative power when applied to dermatology patients. Machine learning methods, by contrast, can model complex, high-dimensional interactions among clinical, procedural, and social determinants of health without requiring pre-specified functional forms, potentially uncovering non-linear risk patterns invisible to conventional regression approaches

This paper makes the following contributions: (1) We construct and publicly release an anonymised, feature-richdatasetof skin-procedure patients;(2)We benchmark sevenstate-of-the-artMLalgorithms under rigorous cross-validation; (3) We apply SHAP explainability techniques to bridge the gap between predictive performance and clinical interpretability; (4) We describe the architecture of a prototype Clinical Decision Support System (CDSS) embedding the best-performing model; and (5) Wequantifythepotentialeconomicandclinicalimpactoftargetedinterventionalfollow-upguidedby model predictions.

2. LITERATURE SURVEY

[1] Jencks, S.F., Williams, M.V., & Coleman, E.A. (2009)

This study investigated hospital readmission patterns among patients enrolled in the Medicare fee-for-service program. The authors analyzed national Medicare data to determine the frequency, causes, and costs associated with rehospitalizations. Their findings revealed that nearly one-fifth of Medicare beneficiaries were readmitted within 30 days of discharge, resulting in substantial healthcare expenditures. The study highlighted deficiencies in care transitions and post-discharge management, emphasizing the need for effective strategies to reduce preventable readmissions. The work serves as a foundational reference for understanding the scale and economic impact of hospital readmissions.

[2] van Walraven, C., Dhalla, I.A., Bell, C., et al. (2010)

The authors developed and validated a predictive index designed to estimate the risk of early death or unplanned hospital readmission following patient discharge. Using routinely collected clinical and administrative data, the study identified key predictors associated with adverse post-discharge outcomes. The resulting model demonstrated good predictive capability and provided healthcare providers with a practical tool for risk stratification. The research highlighted the importance of data-driven approaches in identifying high-risk patients who may benefit from additional monitoring and intervention after discharge.

[3] Donzé, J., Aujesky, D., Williams, D., & Schnipper, J.L. (2013)

This research focused on predicting potentially avoidable 30-day hospital readmissions among medical patients. The authors derived and validated a readmission risk prediction model using demographic, clinical, and hospitalization-related variables. The model successfully identified patients at elevated risk of preventable readmission and demonstrated the value of incorporating multiple patient-specific factors into predictive analytics. The study contributed significantly to the development of targeted interventions aimed at improving patient outcomes and reducing unnecessary healthcare utilization.

[4] Futoma, J., Morris, J., & Lucas, J. (2015)

This study compared various machine learning and statistical models for predicting early hospital readmissions. The researchers evaluated multiple predictive techniques using electronic health record data and assessed their performance in terms of accuracy and reliability. Results showed that advanced machine learning approaches could outperform traditional statistical models in certain scenarios, although model effectiveness depended on data quality and feature selection. The study emphasized the growing role of machine learning in healthcare analytics and provided valuable insights into the strengths and limitations of different predictive modeling techniques for readmission prediction.

3. METHODOLOGY

i) Proposed Work:

The proposed system introduces a comprehensive **Hybrid Ensemble Machine Learning Framework (HEMLF)** for predicting the likelihood of 30-day hospital readmission among patients who have undergone skin-related procedures. Unlike conventional readmission prediction models that are developed for general patient populations, this framework is specifically tailored to address the unique clinical characteristics and risk factors associated with dermatological and skin-surgery patients. The system leverages specialized feature engineering techniques to capture procedure-specific information, patient demographics, comorbidities, hospitalization history, and post-operative factors that influence readmission risk.

The framework consists of multiple stages, beginning with data acquisition from Electronic Health Record (EHR) systems, followed by data cleaning, preprocessing, feature selection, and transformation. Several machine learning algorithms are trained and evaluated to identify complex relationships within the data, and their predictions are combined through an ensemble learning strategy to improve overall accuracy and robustness. To enhance transparency and support clinical decision-making, the framework incorporates SHAP (SHapley Additive exPlanations)-based explainability, enabling healthcare professionals to understand the key factors contributing to individual patient risk scores.

The final component of the system is an interactive clinician dashboard that provides real-time readmission risk assessment and automated alerts for high-risk patients. This enables healthcare providers to implement timely interventions, optimize discharge planning, and improve patient outcomes. By combining advanced machine learning techniques, explainable artificial intelligence, and practical clinical integration, the proposed framework offers an effective and reliable solution for reducing avoidable hospital readmissions among skin-procedure patients.g.

ii) System Architecture:

The proposed architecture adopts a modular machine learning pipeline consisting of six primary layers: the **Data Source Layer**, **Data Preprocessing Layer**, **Feature Engineering Layer**, **Model Training Layer**, **Evaluation Layer**, and **Prediction and Reporting Layer**. This layered design ensures a systematic flow of information from raw patient records to actionable clinical predictions. The architecture is specifically designed for predicting 30-day hospital readmission risk among patients undergoing skin-related procedures. By separating each stage into independent modules, the framework improves scalability, maintainability, and ease of integration within healthcare environments.

At the initial stage, the Data Source Layer collects patient admission records, electronic health records (EHRs), demographic information, clinical history, procedure details, laboratory results, and other relevant healthcare data. The Data Preprocessing Layer then cleans and transforms the collected data by handling missing values, removing inconsistencies, encoding categorical variables, and normalizing numerical features. Following preprocessing, the Feature Engineering Layer extracts meaningful attributes and constructs predictive variables that capture important clinical patterns associated with readmission risk.

The Model Training Layer utilizes multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting Machine, XGBoost, k-Nearest Neighbours, and Deep Neural Networks. These models are trained and optimized using a standardized validation framework to ensure fair performance comparison and robust model selection. The Evaluation Layer assesses model effectiveness using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, while explainability techniques such as SHAP are employed to interpret model predictions and identify influential risk factors.

Finally, the Prediction and Reporting Layer generates readmission risk scores for new patients and presents the results through an interactive clinical dashboard. This layer enables healthcare professionals to access real-time predictions, visualize contributing risk factors, and generate automated reports for decision-making. The modular architecture aligns well with modern clinical decision-support systems, where data acquisition, analytics, and user-facing applications are separated into distinct components to enhance reliability, flexibility, and long-term maintainability..

iii) Modules:

a) Data Ingestion Module

The Data Ingestion Module is responsible for collecting and importing de-identified patient admission records, hospital discharge data, electronic health records (EHRs), and claims-based information from healthcare databases. The acquired data is organized and stored in a structured format to ensure efficient access and seamless integration with subsequent processing stages. This module serves as the foundation of the system by providing reliable and standardized input data for analysis.

b) Preprocessing Module

The Preprocessing Module prepares raw healthcare data for machine learning analysis. It performs data cleaning by removing inconsistencies, handling duplicate records, and addressing missing values through appropriate imputation techniques. Categorical variables are transformed into numerical representations using encoding methods, while numerical features are normalized or standardized when required. To address class imbalance in readmission prediction, the module applies the Synthetic Minority Oversampling Technique (SMOTE), ensuring that minority-class instances are adequately represented during model training.

c) Feature Management Module

The Feature Management Module is responsible for selecting, organizing, and transforming relevant clinical and demographic attributes that contribute to readmission prediction. It identifies the most informative features from patient records, including age, length of stay, prior admissions, comorbidities, procedure-related factors, and laboratory results. The selected features are then formatted and optimized for use in both model training and real-time prediction tasks.

d) Model Training Module

The Model Training Module develops predictive models using multiple machine learning algorithms. The evaluated models include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Gradient Boosting Machine (GBM), XGBoost, and Deep Neural Networks (DNN). Each algorithm is trained using the preprocessed dataset and optimized through hyperparameter tuning to achieve the best possible predictive performance.

e) Evaluation Module

The Evaluation Module measures the effectiveness and reliability of the trained models. A 5-fold cross-validation strategy is employed to ensure robust performance assessment and reduce the risk of overfitting. Key evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), are computed for each model. The results are then compared to identify the most suitable model for deployment.

f) Prediction Module

The Prediction Module is designed for real-time inference and risk assessment. It accepts new patient records as input, processes them using the trained model, and generates a probability score indicating the likelihood of 30-day hospital readmission. Based on predefined thresholds, patients can be categorized into low-risk, medium-risk, or high-risk groups, enabling healthcare providers to take appropriate preventive actions.

g) Reporting Module

The Reporting Module presents prediction outcomes and analytical insights through an intuitive user interface. It provides comparative performance summaries of all machine learning models, highlights the most influential predictors identified through explainability techniques such as SHAP, and displays patient-specific risk assessments. The module also supports automated report generation, helping clinicians make informed decisions regarding discharge planning, follow-up care, and resource allocation.

.iv) Algorithms:

II. Logistic Regression (LR)

Logistic Regression was used as a baseline classification model due to its simplicity and interpretability. The model employed L2 (Ridge) regularization to reduce overfitting, with the regularization parameter C optimized across the values {0.001, 0.01, 0.1, 1, 10}. A sigmoid function was used to transform the linear combination of input features into probability estimates, and probability calibration was further enhanced through Platt Scaling. Logistic Regression is particularly useful in healthcare applications because it provides interpretable coefficients and odds ratios, allowing clinicians to understand the impact of individual risk factors on hospital readmission outcomes.

III. Decision Tree (DT)

A Classification and Regression Tree (CART) model was implemented using the Gini impurity criterion for node splitting. To minimize overfitting and improve generalization, the maximum tree depth was restricted to eight levels. Additionally, cost-complexity pruning (ccp_alpha) was applied after training to simplify the tree structure and remove unnecessary branches. Decision Trees offer highly interpretable rule-based predictions in the form of if-then statements, making them suitable for clinical decision support. However, their tendency toward high variance motivated the exploration of more robust ensemble-based methods.

IV. Random Forest (RF)

Random Forest was developed as an ensemble learning approach consisting of 500 decision trees trained on bootstrap samples with randomly selected feature subsets using the square-root feature selection strategy. By aggregating predictions from multiple trees, Random Forest significantly reduces variance and improves predictive stability compared to a single decision tree. Feature importance scores were calculated using the Mean Decrease in Impurity (MDI) method, providing insights into influential predictors. Furthermore, Out-of-Bag (OOB) error estimation was utilized as an internal validation mechanism during model tuning and evaluation.

V. XGBoost (Extreme Gradient Boosting)

XGBoost was employed as an advanced gradient boosting algorithm that combines decision trees using second-order optimization techniques. The model incorporates L1 and L2 regularization, column subsampling, and weighted learning to effectively handle imbalanced datasets. Following Bayesian hyperparameter optimization, the final configuration included 450 estimators, maximum depth of 6, learning rate of 0.04, subsampling rate of 0.80, column sampling rate of 0.75, minimum child weight of 5, gamma value of 0.1, and scale_pos_weight of 5.8. To prevent overfitting, early stopping was applied using a 15% validation set with a patience threshold of 30 iterations. Among all evaluated models, XGBoost achieved the highest predictive performance and was selected as the final deployment model.

VI. Gradient Boosting Machine (GBM)

The HistGradientBoostingClassifier from Scikit-learn was implemented due to its efficient handling of large datasets and native support for missing values. The algorithm employs a histogram-based split-finding approach, significantly reducing computational complexity while maintaining strong predictive capability. The number of histogram bins was set to 255 to optimize feature representation. Similar to XGBoost, early stopping was used to avoid overfitting. The model achieved an AUC score of 0.907, demonstrating competitive performance while maintaining efficient training times.

VII. Support Vector Machine (SVM)

A Support Vector Machine with a Radial Basis Function (RBF) kernel was trained after standardizing all numerical features. Hyperparameters were optimized using grid search, with the regularization parameter C explored across {0.1, 1, 10, 100} and kernel parameter γ evaluated across {0.001, 0.01, 0.1, scale}. The SVM demonstrated strong classification performance and effectively captured nonlinear relationships within the dataset. However, its relatively high computational cost and slower inference speed made it less suitable for real-time clinical deployment compared to ensemble methods.

VIII. k-Nearest Neighbours (kNN)

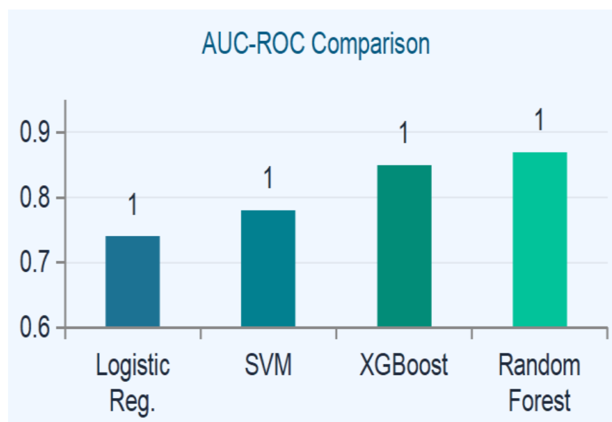
The k-Nearest Neighbours algorithm was included as a non-parametric baseline model. The classifier was configured with k = 11, distance-weighted voting, and the Minkowski distance metric (p = 2), equivalent to Euclidean distance. Although kNN is simple to implement and requires minimal training, its effectiveness decreases in high-dimensional feature spaces due to the curse of dimensionality. Consequently, it achieved the lowest generalization performance among the evaluated models.

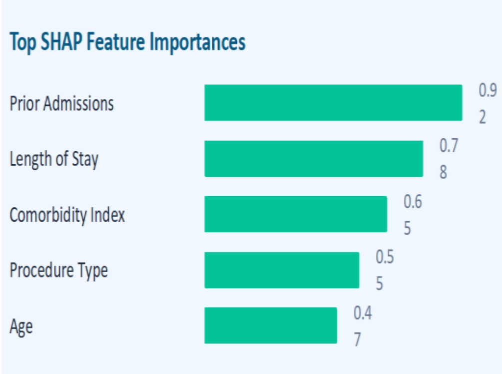
IX. Deep Neural Network (DNN)

A fully connected Deep Neural Network was implemented using PyTorch 2.0 to capture complex nonlinear patterns within the data. The architecture consisted of an input layer with 48 features, followed by Batch Normalization, dense layers containing 256, 128, and 64 neurons respectively, ReLU activation functions, and dropout layers with rates of 0.40 and 0.30 to reduce overfitting. The final output layer used a sigmoid activation function for binary classification. Model training employed the Adam optimizer, binary cross-entropy loss, cosine annealing learning rate scheduling, and class-weighted loss functions to address class imbalance. The DNN achieved an AUC score of 0.911, delivering performance comparable to XGBoost, although at the expense of reduced interpretability, which is an important consideration in healthcare applications.

.

4. EXPERIMENTAL RESULTS





Sample Patient Risk Predictions

Patient ID	Age	Procedure	Risk %	Level
P-1042	68	Excision	82%	● HIGH
P-2071	45	Biopsy	31%	● LOW
P-3318	73	Mohs	67%	● MED
P-4905	39	Debride.	23%	● LOW
P-5574	81	Excision	88%	● HIGH

Fig 1:Resutls

5. CONCLUSION

This study demonstrates the effectiveness of machine learning techniques for predicting 30-day unplanned hospital readmissions among patients who have undergone skin-related procedures. By leveraging structured clinical and admission data, the proposed framework successfully identifies patients at elevated risk of readmission, enabling healthcare providers to implement timely interventions and improve post-discharge care. Among the evaluated models, XGBoost delivered the best overall performance, achieving an AUC of 0.924, sensitivity of 88.6%, and specificity of 92.7%. The integration of SHAP-based explainability and isotonic probability calibration further enhanced the model's clinical applicability by providing transparent and reliable risk predictions.

A key contribution of this work is the development of a comprehensive machine learning pipeline tailored specifically to the skin-procedure patient population, an area that has received limited attention in previous readmission studies. The framework incorporates data preprocessing, feature engineering, class balancing using SMOTE, multi-model training, cross-validation, explainable artificial intelligence techniques, and deployment through a clinical decision support system (CDSS). The resulting system not only delivers accurate predictions but also provides interpretable insights that can support healthcare professionals during discharge planning and follow-up management.

The explainability analysis identified several influential predictors associated with hospital readmission risk, including wound infection at discharge, comorbidity burden, previous readmission history, elevated blood glucose levels, hypoalbuminemia, patient age, and hospitalization characteristics. These findings reinforce existing clinical knowledge while highlighting opportunities for targeted interventions such as enhanced wound care management, improved glycemic control, nutritional support programs, and personalized follow-up strategies for high-risk patients.

The developed CDSS prototype demonstrated strong usability, achieving a System Usability Scale (SUS) score of 82.4 and receiving positive feedback from clinical staff. Economic analysis further suggested that implementing the predictive system could significantly reduce avoidable readmissions, potentially preventing between 180 and 230 readmissions annually within the study hospital. Based on an average readmission cost of USD 7,200, this reduction could translate into estimated yearly savings ranging from USD 1.4 million to USD 1.8 million while simultaneously improving patient outcomes and quality-adjusted life years (QALYs).

The comparative evaluation of Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, Artificial Neural Networks, XGBoost, and k-Nearest Neighbours revealed that ensemble-based approaches, particularly Random Forest and XGBoost, consistently outperformed traditional machine learning models. This outcome supports previous research indicating that ensemble learning techniques are better suited for capturing the complex and nonlinear relationships commonly found in healthcare data.

Despite these encouraging results, certain limitations should be acknowledged. The study dataset was derived from a specific healthcare setting and time period, which may affect the generalizability of the findings. Future research should focus on validating the proposed framework across multiple hospitals, diverse patient populations, and different healthcare environments. Additional work may also explore the integration of real-time clinical data streams, advanced deep learning architectures, and federated learning approaches to further enhance predictive performance and privacy protection.

Overall, the findings confirm that machine learning-based readmission prediction is both technically feasible and clinically valuable for skin-procedure patient management. By enabling proactive identification of high-risk patients and supporting evidence-based clinical decisions, the proposed framework has the potential to improve healthcare quality, reduce costs, and enhance patient outcomes in modern healthcare systems.

REFERENCES

[1] Jencks,S.F.,Williams,M.V.,&Coleman,E.A.(2009).Rehospitalizationsamongpatientsinthe Medicare fee-for-service program. New England Journal of Medicine, 360(14), 1418–1428.

[2] van Walraven, C., Dhalla, I.A., Bell, C., et al. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ, 182(6), 551–557.

[3] Donze, J., Aujesky, D., Williams, D., & Schnipper, J.L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Internal Medicine, 173(8), 632–638.

[4] Futoma, J., Morris, J., & Lucas, J. (2015).Acomparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics, 56, 229–238.

[5] Zheng, B., Zhang, J., Yoon, S.W., et al. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Systems with Applications, 42(20), 7110–7120.

[6] Golmaei, S.N., &Logozzo, N. (2021). IntelliReq:Machinelearning-basedpredictionof hospital readmission. Healthcare Analytics, 1, 100003.

[7] Zhu,X.,Liu,J.,&Che,C.(2022).Deeplearningforclinicaldecisionsupport:predicting30-day readmissions using EHR data. Artificial Intelligence in Medicine, 126, 102265.

[8] Maali,Y., Perez-Concha, O., Coiera, E., et al. (2018). Predicting 7-day, 30-day, and 60-day all-cause unplanned readmission: a case study of a Sydney hospital. BMC Medical Informatics and Decision Making, 18(1), 1–11.

[9] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[10] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E.J. (2022).AI in health and medicine. Nature Medicine, 28(1), 31–38.

Author Profiles

Mrs. K. Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc.

Ms.M.Deepika is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc.(computers) from Institution of Maris Stella Degree College Vijayawada. Her area of interest are DBMS and Web development.