iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55549

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Diabetes Detection System

Likith Shetty[1], Divya Shah[2], Akash Sharma[3], Aditi Gunjal[4], Manya Gidwani[5], Dr. Bhavesh Patel[6]
Shah and Anchor KutchhiEngineering College Chembur, India

Abstract: ***Diabetes is a chronic illness that interferes with your body'sability to control blood sugar or glucose levels. Our body' cells depend on glucose as a major energy source, and the hormone insulin, which thepancreas produces, aids in controlling glucose levels. This research suggests a machine learning based diabetes detection system. The method makes use of a dataset of patient records that detail several clinical traits and whether diabetes is present or not. The most crucial characteristics for predicting diabetes are determined via a feature selection technique. The dataset is then used to train and evaluate machine learning techniques such as Random Forest, Support Vector Machine, KNN and Logistic Regression. The model's performance can be evaluated using certain metrics. The outcomes demonstrate that the suggested system beats the numerous baseline models and has good predictive accuracies for diabetes using Random Forest algorithm. This approach may be helpful for detecting diabetes in early stage and can helpin improving the outcome of patients.***
Index Terms: ***Diabetes, Machine Learning, Diabetes Detection,Dataset, Metrics, Accuracies***

I. INTRODUCTION

Diabetes is a chronic illness that affects millions of individuals worldwideand, if ignored, is linked to serious health issues. Diabetes must be identified and managed early in order to reduce these consequences and enhance patient outcomes. In the field of healthcare, notably in the prediction and diagnosis of diseases, machine learning approaches, particularly the random forest algorithm, have shown considerable promise. In this study, we suggest a random forest-based diabetes diagnosis method.The method makes use of a dataset of patient records that include a numberof clinical characteristics, including age, BMI, blood pressure, and the presence or absence of diabetes. To determine which traits are most crucial for predicting diabetes, we use a feature selection process. The dataset is then used to train and test a random forest model. The random forest algorithm is an ensemble technique that brings together various decision trees to increase prediction accuracy. A randomly chosen subset of featuresand a randomly chosen sample of training data are used to construct each decision tree in the random forest. Combining all of the decision trees' projections yields the ultimate conclusion. The suggested random forest algorithm-based diabetes diagnosis system has a number of advantages over conventional statistical approaches. It can handle highly dimensional, non-linear data and record intricate relationships between various aspects. A personalized approach to interventions and treatments may be made possible by the system's insightson the risk factors for diabetes. The system's effectiveness is assessed using a number of parameters, including accuracy, precision, recall, and F1 score. The outcomes show thatthe suggested method beats numerous baseline models and achieves high accuracy in predicting diabetes. The random forest model's most critical properties for predicting diabetes, according to the feature importance analysis, are age, BMI, and blood pressure.

II. LITERATURE REVIEW

K.VijiyaKumar et al. provides us the information about the Diabetes disease and explains to us how the rapid growth of the disease is affectingthe people. Also they have assumed that by the year 2035 the patients willbe doubled as 592 Million. To detect diabetes in an earlier stage they haveused Random Forest Algorithm as it provides more accurate results than other models[1].

S.Kranthi Reddy et al. have used the PIMA Indian Diabetes Dataset from Kaggle to detect the diabetes in female patients. They have used two algorithms namely Random Forest and K-Nearest Neighbor having accuracy of 78.4% and 80.8% respectively. Where they used K-Fold Crossvalidation technique to train the dataset applying 11 folds [2].

L.V.Rajani Kumari et al. provides us information about how diabetes is affecting people and how Machine Learning can be used to predict diabetesin early stages. They have used several Machine Learning models such asK-Nearest Neighbor, Logistic Regression, Naive Bayes and Random Forest having accuracy about 78.57%, 72%, 71% and 76% respectively. By using Evaluation Metrics they have explained which model is best suited for predicting diabetes [10].

Deeraj Shetty et al. explains how Data mining techniques can be used to extract useful data from the Hospital data. The primary target of this examination is to assemble an Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's databases.

In this system, we propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on a diabetes patient's database and analyze them by taking various attributes of diabetes for prediction ofdiabetes disease [3].
Narendra Mohan et al. explains how support vector machine [SVM] is applied in diabetes prediction. The performance of the SVM algorithm is analyzed for different available kernels. The best kernel is selected and used for prediction. The proposed work is implemented in python programming language and its performance is as good as other algorithms[5].

III. PROPOSED SYSTEM

A. System Architecture of Diabetes Detection System.

The proposed System Architecture of Diabetes Detection System is describedin the figure below:
In this project, the major language used is Python with the help of Jupyter Notebook and Streamlit for creating the front-end. The Oversampling Technique SMOTE is used to balance the dataset while K-Fold Cross Validation Technique is used to train the Dataset. With the application of Random Forest Algorithm the accuracy score of 83% has been achieved. Wetransformed our model into a pickle model to connect it with the front-end. Hence, we created a web application to detect Diabetes in person using the medical parameters and values.

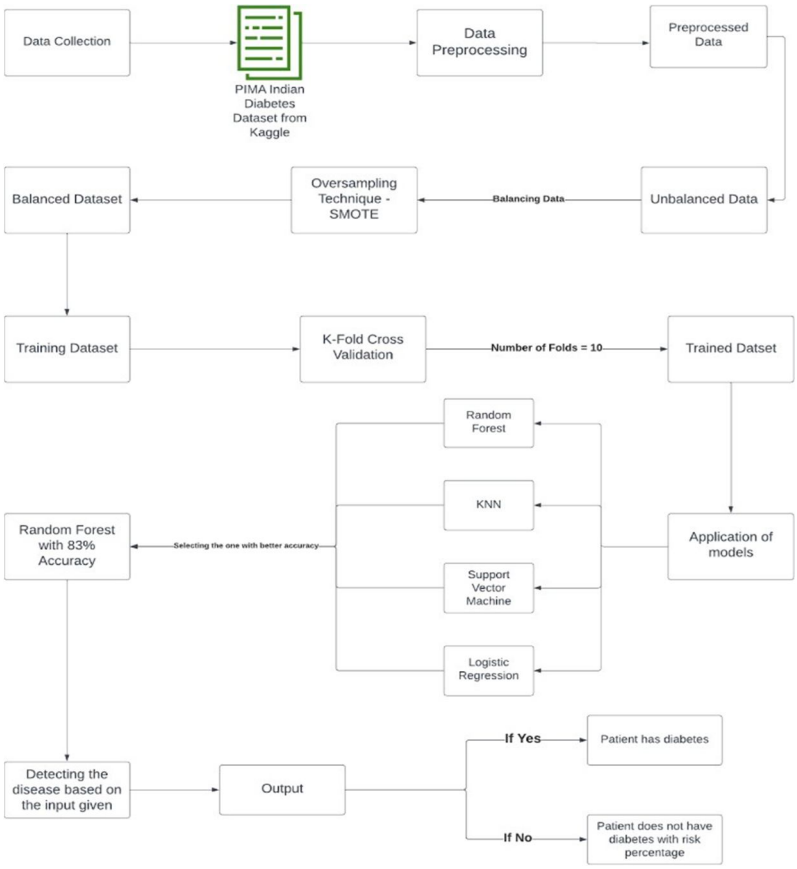


Fig: System Architecture

B. Implementation

The first step is to collect the dataset which we did from Kaggle. The name ofthe dataset is PIMA Indian Diabetes Dataset. This dataset consists of female patients with a minimum age of 21 consisting of their diagnostic parameters. After this step, the next step was data preprocessing which includes data cleaning and removal of null values. Now it is seen that the dataset has 768 values which includes 500 data points for Non-Diabetic patients and 268 datapoints for Diabetic patients. So it is evident that the dataset is highly imbalanced which can cause issues for the models that have to be applied. Hence, to solve these issues a balancing technique can be used such as oversampling.

Oversampling is a technique used in machine learning to balance the imbalanced dataset by increasing the instances of minority class. So here we used an Oversampling Technique called SMOTE(Synthetic Minority Over Sampling Technique). SMOTE generates synthetic samples ofminority class by interpolating between existing samples.

```
0    500                    1    500
1    268                    0    500
Name: Outcome, dtype: int64   Name: Outcome, dtype: int64
```

Fig: Before Smote Fig: After Smote

After balancing the dataset, the K-Fold Cross Validation method is used to train and evaluate the model's performance. It involves dividing the dataset into k subsets or folds of roughly equal size and then training the model on k-1 folds and using the remaining fold as a validation set to evaluate the model's performance. Evaluation metrics such as precision, recall, F1 score are used to evaluate the model's performance.

Model	Accuracy	Precision	Recall	F1 Score	Support
Random Forest	0.8300	0.8018	0.8812	0.8396	200
Model	Accuracy	Precision	Recall	F1 Score	Support
Logistic Regression	0.7500	0.7476	0.7624	0.7549	200
Model	Accuracy	Precision	Recall	F1 Score	Support
SVC	0.7450	0.7451	0.7525	0.7488	200
Model	Accuracy	Precision	Recall	F1 Score	Support
KNN	0.7300	0.7009	0.8119	0.7523	200

Fig: Evaluation Metrics

After all the evaluation it is evident that the Random forest Algorithm has better evaluation metric scores than other model's. Random Forest having Accuracy Score : 83%, Precision : 80.18%, Recall : 88.12% and F1 Score :83.96%.

```
Cross Validation accuracies for  RandomForestClassifier(random_state=42) =  [0.85 0.86 0.88 0.78 0.79 0.77 0.83 0.88 0.86 0.81]
Accuracy % of the  RandomForestClassifier(random_state=42) 83.0
[[77 22]
 [12 89]]
```

Fig: Confusion matrix of Random forest model

Now we made a ROC curve for Random forest. A Receiver Operating Characteristic (ROC) is a graphical representation of a classification model. It is a plot of True Positive Rate (TPR) against False Positive Rate (FPR). The model having AUC value above 0.5 is considered to be a better model for the given dataset while the model having AUC value below 0.5 is considered to bea good model for given dataset. Here Random Forest model gives AUC value
= 0.89 and hence can be considered as a great option to opt.

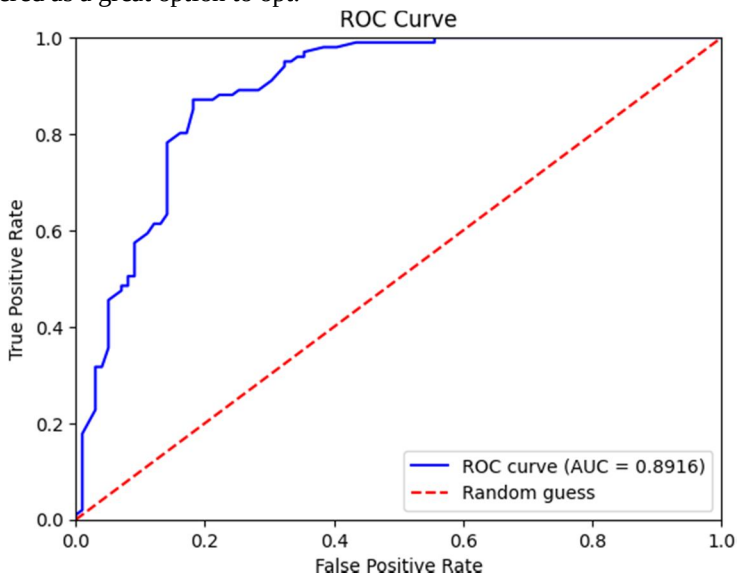


Fig: ROC Curve for Random Forest

Now after the application of K-Fold Cross Validation we have broken down the probabilities 0 and 1 to have a clear understanding of the results as predicted by Random Forest Algorithm. So it is found that data points that have probability under 0.5 are considered to be Non-diabetic patients while data points that have probability over 0.5 are considered to be Diabetic patients.

```
for model in models:
    prob = model.predict_proba(X_resampled)
print(prob)
```

```
[[0.03 0.97]
 [0.88 0.12]
 [0.16 0.84]
 ...
 [0.   1.  ]
 [0.27 0.73]
 [0.02 0.98]]
```

Fig. Broken down probabilities into 2D arrays.

Then by using Pickle we dumped the model of diabetes and then used it for designing the front end using Streamlit. Now using the broken down values we calculated the risk percentage of the Non-Diabetic Patients to predict the risk of Non-Diabetic patients to develop Diabetes in future.

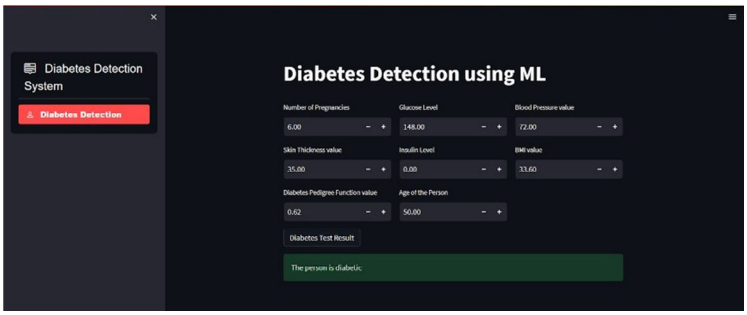


Fig. Prediction for Diabetic Patients.

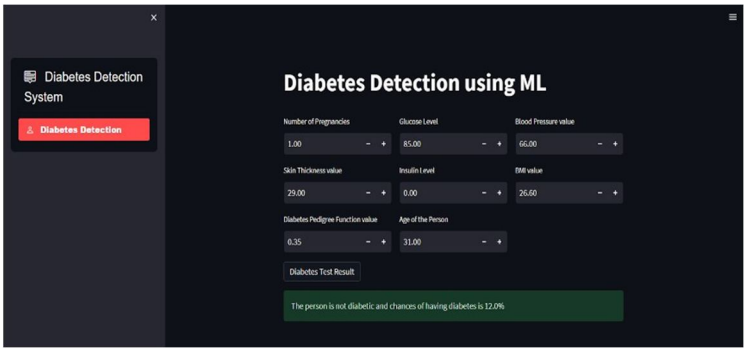


Fig. Prediction for Non-Diabetic Patients.

IV. FUTURE SCOPE

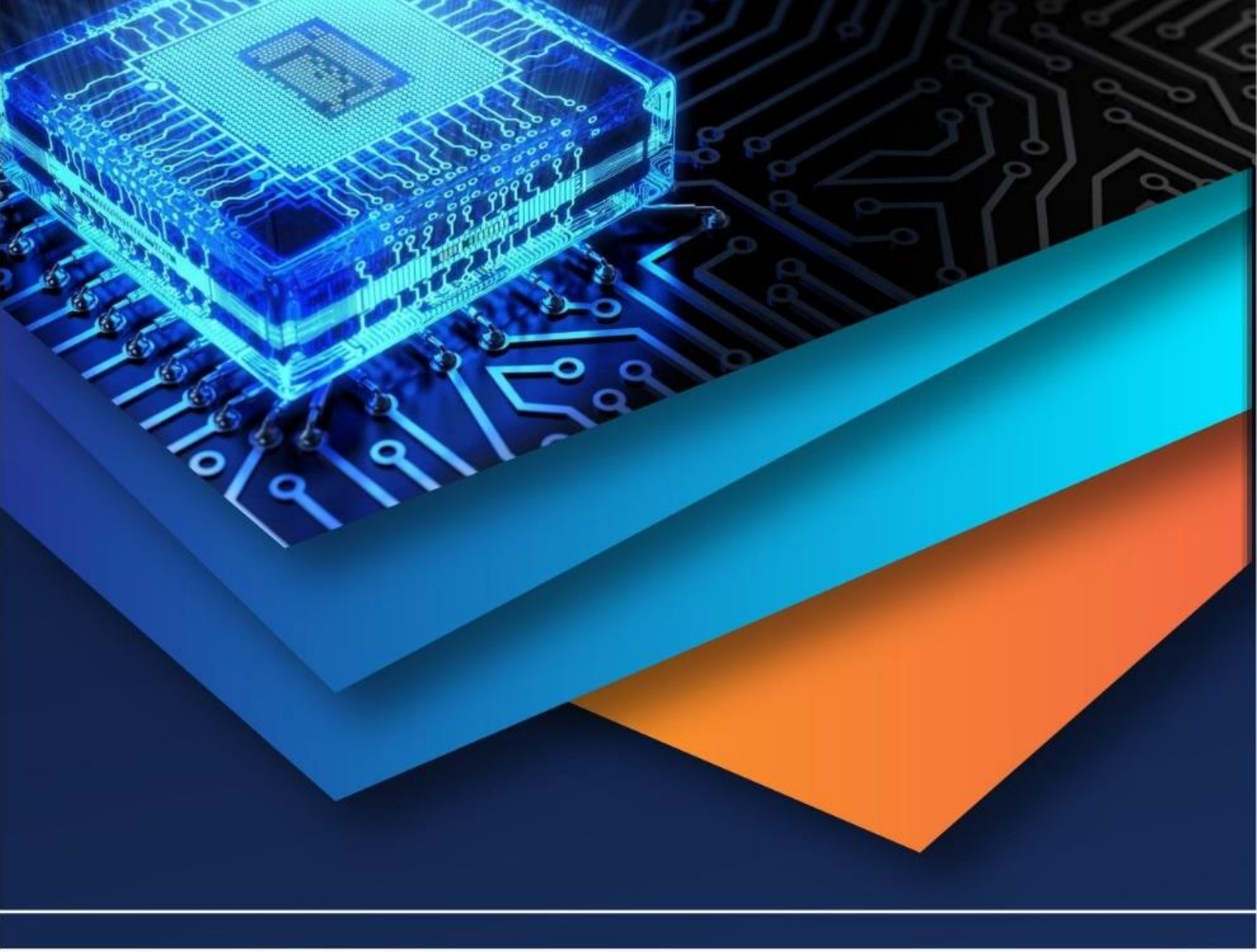
1) *Adding New Features:* To increase the detection of diabetes, the system can be improved by adding new clinical elements, such as genetic data and lifestyle factors.
2) *Utilizing Variable Technologies:* The system can be improved by integrating wearable technology to offer continuous data on patient health, such as that provided by smartwatches and activity trackers.
3) *Real Time Monitoring:* Clinical support systems can be integrated into the system to give healthcare professionals real-time forecasts and insights into the variables influencing diabetes risk.
4) *Personalized Interventions:* Interventions that are specifically tailored to the needs of each patient can be treated using the system, which will enhance patient outcomes and lower healthcare expenditures.
5) *Extending to Other Diseases:* Other diseases can be added to the system to help with early identification and care, such as cancer and cardiovascular disease.

V. CONCLUSION

The proposed system is an effective instrument for the early diagnosis and management of diabetes. In terms of forecasting diabetes, the Random Forest algorithm performed better than numerous other baseline models. By implementing more sophisticated machine learning methods and interfacing with the electronic health record systems, the system can be improved. The technology has the potential to fundamentally alter how medical professionals diagnose and manage diabetes, resulting in better patient outcomes and higherquality of life for diabetic patients.

REFERENCES

[1] K.Vijayakumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, " Random Forest Algorithm for Prediction of Diabetes", International Conference on Systems Computation and Networking, 2019.
[2] S Kranthi Reddy, T Krishnaveni, G.Nikitha, E.Vijayakanth, " Diabetes Prediction Using Different Machine Learning Algorithms ", Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021).
[3] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, " Diabetes Disease Prediction Using Data Mining ", 2017 International Conference on Innovations in Information, Embedded and Communication System (ICIIECS).
[4] Liu Lei, " Prediction of Score of Diabetes Progression Index Based on Logistic Regression Algorithm", 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS).
[5] Narendra Mohan, Vinod Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction", Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020).
[6] Prof.Rajesh Lomte, Sheetal Dagale, Sneha Bhosale, Shraddha Ghodake, "Survey Of Different Feature Selection Algorithms For Diabetes Mellitus Prediction", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
[7] Hala Alshamlan, Hind Bin Taleb, Areej Al Sahow, " A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression ", 2020 11th International Conference on Information and Communication Systems (ICICS).
[8] Vinod Jain, "Performance Analysis of Supervised Machine Learning Algorithm for Prediction of Diabetes", International Conference on Edge Computing and Applications (ICECAA 2022).
[9] Prakhar Saxena, Subhadeep Saha, S. Kiruthika Devi, " Analysis and Prediction of Diabetes Using Machine Models", 2022 International Mobile and Embedded Technology Conference (MECON).
[10] Mrs. L.V Rajani Kumari, P. Shreya, Mehrunnisa Begum, T. Pavan Krishna, M. Prathibha, " Machine Learning Based Diabetes Detection", 2021 6th International Conference On Communication And Electronic Systems (ICCES).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY