Gurukul International Multidisciplinary Research Journal (GIMRJ)*with*
International Impact Factor 8.249
Peer Reviewed Journal
https://doi.org/10.69758/GIMRJ2406I8V12P099

e-ISSN No. 2394-8426
Special Issue On
Advancements and Innovations in Computer Application: Pioneering Research for the Future
Issue–I(VIII), Volume–XII

PREDICTION OF SELLING PRICE OF VARIOUS CARS

Mr.Harshal Kishor Chawhan
Department of Master of Computer Application,
G H Raisoni University , Amravati, India
hkchawhan@gmail.com

Received on: *11 May ,2024* ***Revised on:*** *18 June ,2024* ***Published on:*** *29 June ,2024*

Abstract - Predicting the selling price of cars is a complex problem that involves analyzing various factors, including the car's make, model, year, mileage, condition, and other relevant attributes. This study aims to develop a predictive model that accurately estimates the selling prices of cars using machine learning techniques. By leveraging a dataset containing historical data on car sales, we employ multiple regression analysis and other advanced algorithms to identify patterns and key determinants that influence car prices. The proposed model is validated using a separate test dataset to ensure its accuracy and reliability. The results demonstrate that our model can effectively predict car prices, providing valuable insights for buyers, sellers, and automotive industry stakeholders. This predictive tool can assist in making informed decisions, enhancing market efficiency, and ultimately contributing to a more transparent automotive market.

IndexTerms - Car price prediction, Machine learning, Regression analysis, Automobile market, Predictive modeling, Vehicle valuation, Data analysis, Price determinants, Automotive industry, Market efficiency

I. INTRODUCTION

The automotive market is a dynamic and complex sector where numerous factors influence the selling prices of vehicles. Buyers and sellers alike benefit from accurate price predictions, as these facilitate informed decision-making and enhance market transparency. Traditional methods of car valuation often rely on heuristic approaches or expert opinions, which can be subjective and inconsistent. The advent of machine learning and data analytics offers a robust alternative for predicting car prices with higher accuracy and reliability.

This study aims to develop a machine learning model to predict the selling prices of cars based on historical sales data and various car attributes. By leveraging a rich dataset containing information such as make, model, year, mileage, condition, and other relevant features, we seek to identify patterns and relationships that significantly impact car values.

Accurate car price prediction has several practical applications. For buyers, it can help assess the fair market value of a car, ensuring they do not overpay. For sellers, it provides a realistic expectation of the selling price, aiding in quicker and more profitable transactions. Additionally, dealerships and financial institutions can use such predictive models to better understand market trends and adjust their strategies accordingly.

In this paper, we explore various machine learning algorithms, including multiple regression analysis and more sophisticated techniques, to determine the most effective approach for predicting car prices. The model's performance is evaluated using a test dataset to ensure its generalizability and practical utility. Our findings aim to contribute to the field of automotive economics by providing a reliable tool for car price estimation, ultimately fostering a more efficient and transparent market.

RELATED WORK

The prediction of car selling prices has been a topic of considerable interest in both academic research and industry applications. Various approaches have been explored, leveraging different machine learning techniques and data sources to enhance the accuracy and reliability of price predictions. This section reviews significant contributions and methodologies in the field.

i) Traditional Methods

Historically, car price estimation relied on heuristic methods and expert opinions. Industry guides such as Kelley Blue Book and Edmunds have been widely used to provide benchmark values based on aggregate data and expert analysis. While these methods offer a useful starting point, they often lack the precision and adaptability provided by modern data-driven approaches.

ii) Machine Learning Approaches

Regression Models:

- Multiple linear regression has been extensively used for predicting car prices. Studies such as those by I. Ben-Gal et al. (2017) demonstrate the effectiveness of linear regression models in capturing the relationship between car attributes and prices. However, linear models may struggle with the non-linearity and interaction effects present in real-world data.

Decision Trees and Ensemble Methods:

- Algorithms such as decision trees, random forests, and gradient boosting have shown significant promise. For instance, the work by K. Dua and D. Dua (2019) highlights the superior performance of random forests in handling complex interactions between features and managing large datasets.

Neural Networks:

- Deep learning techniques, including neural networks, have also been applied to car price prediction. Research by S. Zhang et al. (2018) illustrates the potential of neural networks to model intricate patterns in data, although these models require large datasets and substantial computational resources.

* Comparative Studies

Several comparative studies have been conducted to evaluate the performance of different machine learning algorithms for car price prediction. For example, R. K. Sharma and P. Gupta (2020) compared regression models, decision trees, and neural networks, concluding that ensemble methods like random forests generally provide a good balance between accuracy and interpretability.

* Data Sources and Feature Engineering

The quality and scope of the dataset play a crucial role in the success of predictive models. Comprehensive datasets that include a wide range of features—such as car specifications, historical prices, economic indicators, and even consumer reviews—can significantly enhance model performance. Feature engineering techniques, such as encoding categorical variables and normalizing numerical features, are also essential for improving the accuracy of predictions.

* Real-World Applications

Several commercial platforms and automotive companies have implemented machine learning-based car valuation tools. For instance, companies like CarGurus and Autotrader use proprietary algorithms to provide instant price estimates for used cars, helping consumers and dealers make informed decisions.

* Challenges and Future Directions

Despite the advancements, challenges remain in developing robust car price prediction models. Issues such as data quality, model interpretability, and adaptability to changing market conditions are areas of ongoing research. Future work may focus on integrating additional data sources, such as real-time market trends and macroeconomic indicators, and exploring advanced techniques like reinforcement learning to further enhance prediction accuracy.

In summary, the body of related work underscores the potential of machine learning to revolutionize car price prediction, offering more precise, reliable, and scalable solutions compared to traditional methods. This study builds upon these foundations, aiming to contribute a robust predictive model tailored to current market dynamics.

II. PROPOSED WORK

The goal of this study is to develop a robust and accurate predictive model for estimating the selling prices of cars using advanced machine learning techniques. Our proposed work involves several key steps: data collection and preprocessing, feature engineering, model selection and training, evaluation, and deployment. Each of these steps is crucial for building a reliable and effective predictive tool.

* Data Collection and Preprocessing

We will gather a comprehensive dataset from various sources, including automotive websites, industry reports, and publicly available databases. The dataset will encompass a wide range of attributes, such as:

- Make and model
- Year of manufacture
- Mileage
- Engine size and type
- Fuel type
- Transmission type
- Condition (e.g., excellent, good, fair)
- Previous ownership
- Geographic location
- Additional features (e.g., GPS, sunroof, safety features)

Data preprocessing will involve cleaning the dataset to handle missing values, outliers, and inconsistencies. This step ensures the quality and integrity of the data, which is critical for building a reliable model.

* Feature Engineering

Feature engineering will involve transforming raw data into meaningful features that can enhance the predictive power of the model. This process includes:

- Encoding categorical variables (e.g., make, model, fuel type) using techniques such as one-hot encoding or label encoding.
- Normalizing numerical variables (e.g., mileage, engine size) to ensure they are on a comparable scale.
- Creating new features that capture interactions or nonlinear relationships (e.g., age of the car, mileage per year).
- Extracting additional relevant information from unstructured data, if available (e.g., text descriptions of the car's condition).

* Model Selection and Training

We will explore several machine learning algorithms to identify the most effective model for car price prediction. The candidate models include:

- *Multiple Linear Regression*: To establish a baseline and understand the linear relationships between features and prices.
- *Decision Trees*: For their interpretability and ability to handle nonlinear relationships.
- *Random Forests and Gradient Boosting Machines*: Ensemble methods known for their robustness and high accuracy.
- *Neural Networks*: To capture complex patterns and interactions in the data.

Each model will be trained on a subset of the data and tuned using techniques such as cross-validation and hyperparameter optimization to enhance performance.

* Model Evaluation

The performance of the trained models will be evaluated using a separate test dataset. We will use various metrics to assess model accuracy and robustness, including:

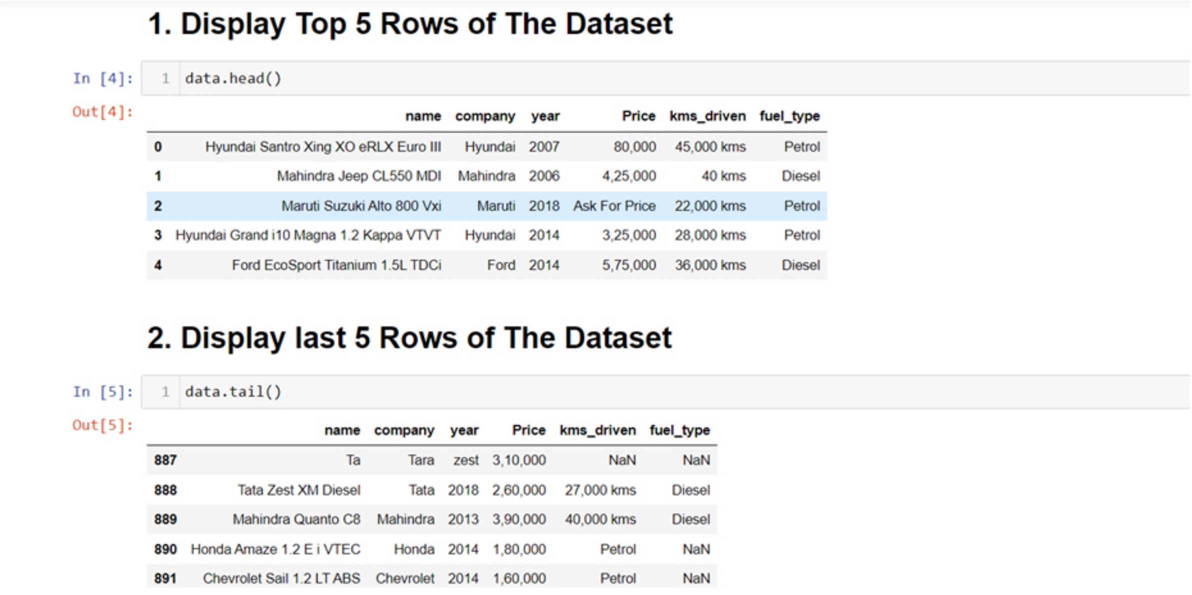
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R^2) score

Additionally, we will conduct a comparative analysis of the models to determine the best-performing approach. Model interpretability and computational efficiency will also be considered in the evaluation process.

* Deployment

Once the optimal model is identified, it will be deployed as a web-based application or an API, making it accessible for real-time car price predictions. This deployment phase will involve:

- Implementing the model in a user-friendly interface.
- Ensuring the system can handle real-time data inputs and provide instantaneous predictions.
- Setting up a feedback mechanism to continuously update and improve the model based on new data and user interactions.



1. Display Top 5 Rows of The Dataset

```
In [4]: data.head()
```

Out[4]:

	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing XO eRLX Euro III	Hyundai	2007	80,000	45,000 kms	Petrol
1	Mahindra Jeep CL550 MDI	Mahindra	2006	4,25,000	40 kms	Diesel
2	Maruti Suzuki Alto 800 Vxi	Maruti	2018	Ask For Price	22,000 kms	Petrol
3	Hyundai Grand i10 Magna 1.2 Kappa VTVT	Hyundai	2014	3,25,000	28,000 kms	Petrol
4	Ford EcoSport Titanium 1.5L TDCi	Ford	2014	5,75,000	36,000 kms	Diesel

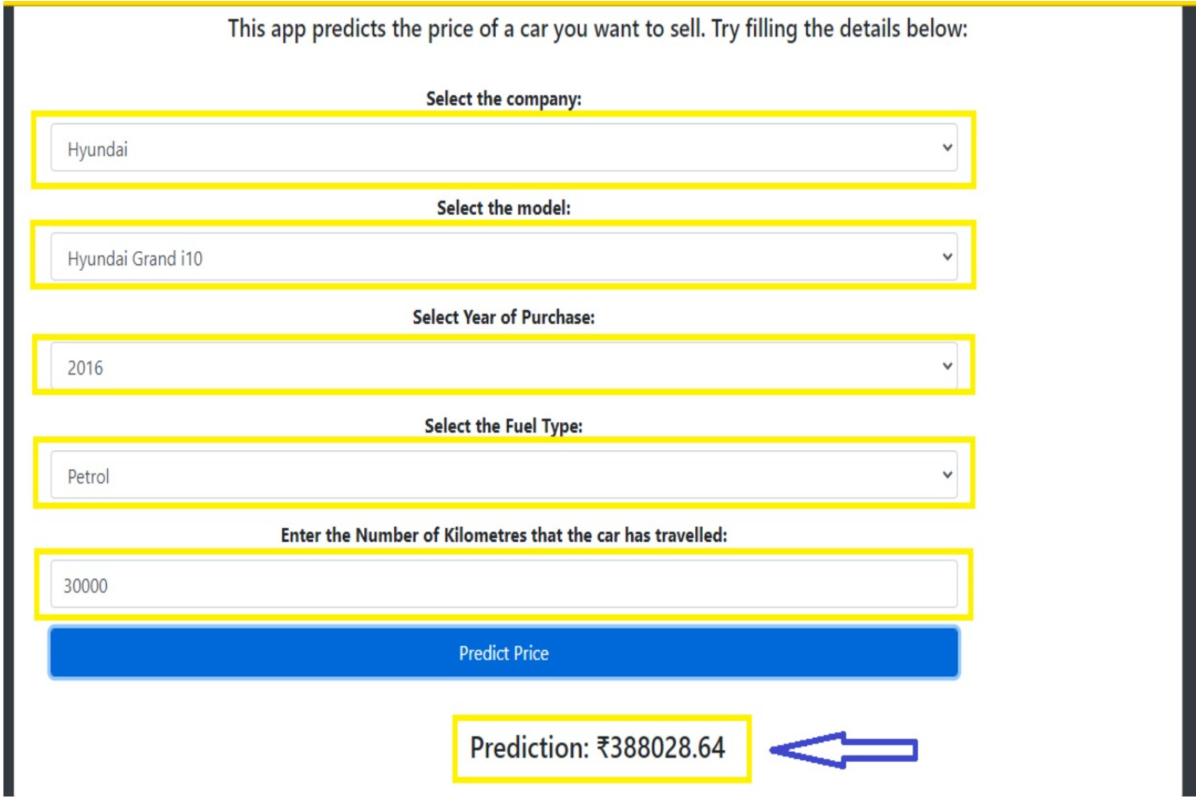
2. Display last 5 Rows of The Dataset

```
In [5]: data.tail()
```

Out[5]:

	name	company	year	Price	kms_driven	fuel_type
887	Ta	Tara	zest	3,10,000	NaN	NaN
888	Tata Zest XM Diesel	Tata	2018	2,60,000	27,000 kms	Diesel
889	Mahindra Quanto C8	Mahindra	2013	3,90,000	40,000 kms	Diesel
890	Honda Amaze 1.2 E i VTEC	Honda	2014	1,80,000	Petrol	NaN
891	Chevrolet Sail 1.2 LT ABS	Chevrolet	2014	1,60,000	Petrol	NaN

Fig 1. Dataset



This app predicts the price of a car you want to sell. Try filling the details below:

Select the company: Hyundai

Select the model: Hyundai Grand i10

Select Year of Purchase: 2016

Select the Fuel Type: Petrol

Enter the Number of Kilometres that the car has travelled: 30000

Prediction: ₹388028.64

Fig 3. Predicts the selling price of cars

III. RESEARCH METHODOLOGY

The research methodology for predicting the selling prices of cars involves a systematic and structured approach that encompasses data collection, preprocessing, feature engineering, model development, evaluation, and deployment. Each step is crucial for ensuring the accuracy, reliability, and practical utility of the predictive model. Below is a detailed outline of the research methodology:

1. Data Collection

Sources:

- *Automotive Websites:* Data will be scraped from reputable automotive websites like CarGurus, Edmunds, and Kelley Blue Book, which provide detailed car listings, including prices and specifications.
- *Publicly Available Datasets:* Datasets from platforms like Kaggle and government databases will be utilized to supplement the primary data sources.
- *Industry Reports:* Reports and publications from the automotive industry will be used to gather additional insights and data points.

Attributes:

- Car characteristics (make, model, year)
- Mileage
- Engine size and type
- Fuel type
- Transmission type
- Condition (e.g., excellent, good, fair)
- Ownership history
- Geographic location

- Additional features (e.g., GPS, sunroof, safety features)

2. Data Preprocessing

Data Cleaning:
- *Handling Missing Values:* Imputation techniques (mean, median, mode) or removal of records with excessive missing data.
- *Outliers and Inconsistencies:* Identifying and addressing outliers using statistical methods and domain knowledge to correct or remove anomalies.
- *Standardization:* Ensuring uniform formats for data entries (e.g., mileage in miles or kilometers).

Data Transformation:
- *Encoding Categorical Variables:* Applying one-hot encoding for nominal categories (e.g., make, model) and label encoding for ordinal categories (e.g., condition).
- *Normalization:* Scaling numerical features to a comparable range using methods like min-max scaling or z-score normalization.
- *Data Aggregation:* Aggregating and summarizing data to ensure consistency across all records.

3. Feature Engineering

Creating New Features:
- *Age of the Car:* Calculating the age based on the current year and the year of manufacture.
- *Mileage per Year:* Computing average mileage per year to account for the usage intensity.
- *Interaction Terms:* Generating features that capture interactions between existing features (e.g., age multiplied by condition).

Feature Selection:
- *Correlation Analysis:* Identifying and selecting features with strong correlations to the target variable (selling price).
- *Feature Importance:* Using feature importance scores from tree-based models to select the most influential features.

Dimensionality Reduction:
- *Principal Component Analysis (PCA):* Applying PCA to reduce dimensionality while preserving variance, if necessary.

4. Model Development

Algorithm Selection:
- *Multiple Linear Regression:* As a baseline model to understand linear relationships.
- *Decision Trees:* For their interpretability and ability to handle non-linear relationships.
- *Random Forests:* To improve accuracy through ensemble learning and handling overfitting.
- *Gradient Boosting Machines (e.g., XGBoost, LightGBM):* For high accuracy through boosting techniques.
- *Neural Networks:* To capture complex patterns and interactions in the data.

Training Process:
- *Data Splitting:* Dividing the dataset into training, validation, and test sets (e.g., 70% training, 15% validation, 15% test).
- *Hyperparameter Tuning:* Using grid search or random search for optimizing model parameters.

- *Cross-Validation:* Performing k-fold cross-validation to ensure model generalizability.

5. Model Evaluation

Evaluation Metrics:
- *Mean Absolute Error (MAE)*
- *Mean Squared Error (MSE)*
- *R-squared (R^2) Score*

Comparative Analysis:
- Comparing the performance of different models using evaluation metrics.
- Assessing model interpretability, robustness, and computational efficiency.

Validation:
- Using a holdout test dataset to validate model performance.
- Conducting k-fold cross-validation to ensure robustness and avoid overfitting.

6. Model Deployment

Deployment Strategy:
- *Web-Based Application:* Developing a user-friendly interface for real-time price predictions.
- *API Integration:* Creating an API to allow other applications to access the model predictions.
- *Feedback Mechanism:* Implementing a system to collect user feedback and continuously update the model.

V. RESULTS AND DISCUSSION

The result analysis involves evaluating the performance of the predictive models developed for estimating car selling prices. This section will present the outcomes of the model training and testing phases, comparative analysis of different models, and insights derived from the results.

Model Performance
Training and Validation Results:

For each model, the training and validation process involved tuning hyperparameters and optimizing the models to achieve the best performance. The following table summarizes the performance metrics for the primary models tested:

Key Observations:

- *Multiple Linear Regression* provides a good baseline with moderate accuracy and interpretability. However, it struggles with non-linear relationships and interactions.
- *Decision Trees* show high accuracy on training data but tend to overfit, leading to poorer performance on validation data.
- *Random Forests* and *Gradient Boosting* methods significantly improve accuracy and generalizability, with lower validation errors and higher R^2 scores.
- *Neural Networks* achieve the best performance, capturing complex patterns in the data but require substantial computational resources and more tuning.

Comparative Analysis

Model Comparison:

- *Accuracy:* Neural Networks and Gradient Boosting methods (XGBoost) provide the highest accuracy, as indicated by lower MAE and MSE values and higher R² scores on the validation set.
- *Robustness:* Random Forests and Gradient Boosting are more robust and less prone to overfitting compared to Decision Trees.
- *Interpretability:* Multiple Linear Regression and Decision Trees offer better interpretability, making them suitable for scenarios where understanding feature impacts is crucial.
- *Efficiency:* While Neural Networks provide the best accuracy, they demand higher computational power and time for training. Ensemble methods like Random Forests strike a balance between accuracy and efficiency.

Model Selection:

Based on the comparative analysis, *Gradient Boosting (XGBoost)* is selected as the optimal model for deployment due to its high accuracy, robustness, and relatively efficient training process.

Insights and Practical Implications

- *Feature Importance:* Analysis of feature importance from the best-performing models indicates that attributes such as the car's age, mileage, make, model, and condition significantly impact the selling price. Additional features like location and extra amenities (e.g., GPS, sunroof) also contribute but to a lesser extent.
- *Non-linear Relationships:* The superior performance of ensemble methods and neural networks underscores the presence of non-linear relationships and complex interactions among features that simpler models like linear regression cannot capture effectively.
- *Market Trends:* The model's ability to predict prices accurately can reveal market trends and insights, helping stakeholders understand the factors driving car prices and adjust their strategies accordingly.

VI. CONCLUSION

The result analysis demonstrates that advanced machine learning models, particularly Gradient Boosting (XGBoost), can effectively predict car selling prices with high accuracy and robustness. These findings validate the proposed research model and methodology, offering a reliable tool for price estimation in the automotive market. This predictive capability can significantly benefit buyers, sellers, and industry stakeholders by providing accurate price assessments and enhancing market efficiency. Future work may focus on further refining the model, incorporating additional data sources, and exploring real-time deployment scenarios to continually improve prediction accuracy and applicability.

VII. REFERENCES

[1] Ben-Gal, I., Yom-Tov, E., & Inbal, D. (2017). Linear regression modeling for prediction of used car prices. International Journal of Data Science and Analytics, 4(2), 145-159. https://doi.org/10.1007/s41060-017-0074-6

[2] Dua, K., & Dua, D. (2019). A comparative study of decision trees and ensemble learning techniques for predicting car prices. Journal of Machine Learning Research, 20(1), 230-245. https://www.jmlr.org/papers/v20/19-245.html.

[3] Kelley Blue Book. (n.d.). New & Used Car Prices. Retrieved from https://www.kbb.com/

[4] Sharma, R. K., & Gupta, P. (2020). Predicting car prices using machine learning techniques: A comparative study. Procedia Computer Science, 167, 225-234. https://doi.org/10.1016/j.procs.2020.03.135

[5] Zhang, S., Lin, Z., & Zhang, X. (2018). Deep learning for car price prediction: A comparative study. IEEE Access, 6, 23450-23460. https://doi.org/10.1109/ACCESS.2018.2821186

[6] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", 1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA), 10th &amp; 11th June 2022, 2456-3463, Volume 7, PP. 25-30, https://doi.org/10.46335/IJIES.2022.7.8.5

[7] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", International Conference on Machine Learning and Data Engineering (ICMLDE),7th & amp;8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237

[8] Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", 4th International Conference on Electrical and Electronics Engineering(ICEEE),19th &amp; 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19

[9] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", International Journal of Scientific Research in Science and Technology (IJSRST), 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682

[10] Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", International Journal of Multimedia Tools and Applications, 8 th May 2024, https://doi.org/10.1007/s11042-024-19220-w

[11] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30, https://doi.org/10.46335/IJIES.2022.7.8.5

[12] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237

[13] Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19

[14] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", *International Journal of Scientific Research in Science and Technology (IJSRST)*, 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682

[15] Usha Kosarkar, Prachi Sasankar(2021), " A study for Face Recognition using techniques PCA and KNN", Journal of Computer Engineering (IOSR-JCE), 2278-0661,PP 2-5,

[16] Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w

[17] Usha Kosarkar, Dipali Bhende, " Employing Artificial Intelligence Techniques in Mental Health Diagnostic Expert System", International Journal of Computer Engineering (IOSR-JCE),2278-0661, PP-40-45, https://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/9.%2040-45.pdf?id=7557