Beyond Parental Height: A Multi-Model Deep Learning Approach for Personalized Adult Height Prediction

S.Siva Nageswara Rao
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
profssnr@gmail.com

Siva Nagendra Akurathi
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
sivanagendra2004@gmail.com

Guntur Gowtham
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
gowthamguntur25@gmail.com

Challa Ravi Sankar
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
ravichalla023@gmail.com

Marella Venkata Rao
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
venkatmarella670@gmail.com

***Abstract*—A new multi-model deep learning approach is proposed for the prediction of adult height from the Galton historical dataset with advanced feature engineering. Traditional approaches to height prediction have relied on linear relationships between the heights of parents and their offspring, which cannot explain the more intricate interplay of genetic, environmental, and lifestyle factors. This study bridges the gap by integrating other influences, such as birth order and physical activity, in developing a more holistic model for height prediction. Experiments were performed on raw and processed data, mainly on the impact of removing outliers on the accuracy of the model. Results reflect the fact that a multi-modeling system would predict better than the single-model scheme because combining all the factors is thought to make it more flexible and reliable. Data pre-processing was a very important activity, particularly outlier handling since results indicated that the predictive accuracy significantly improved when outliers were removed. This underlines robust data cleaning in machine learning algorithms. In summary, this study furnishes the pediatrician and the parent with a useful tool in delivering more reliable growth forecasts: underlining the role advanced data science techniques can play within personalized healthcare. This advances height prediction but lays grounds for further studies in individually modeling growth.**

***Index Terms*—Child's adult height prediction (AHP), data analysis, machine learning, healthcare.**

I. Introduction

One of the health metrics is being tall, which is maintained by the interplay of complex genetic and environmental factors. In this light, WHO, CDC, among others, monitor children's growth by tracking data associated with population heights. Cohort studies have enlightened the way growth rates correlate [1]. A major dataset involved in growth curve research and models, such as the Quadratic Exponential Pubertal Halt Model, consists of Galton's height data. Among all the predictors, adult height is one of the critical predictors that determine the performance of athletes in the sport [2]. The major predictors are chronological age, sitting height, and leg length.

Machine learning algorithms have over the last few years gained attention as potent ways of predicting adult height. Some researchers have used Galton's data set to illustrate advancements in predictive efficiency by using such algorithms [3]. Parental height is a relation that would encompass genetic and environmental effects [4]; we used only Galton's data for that. We show the relationship of parent to child height through regression using multimodel deep learning that includes feature engineering as well as outlier removal in the light of further improving prediction of adult heights-a great potential as a resource for pediatricians and parents.

II. Related Work

Recently, good research interests have been drawn on the predictability of adult height using the technique of machine learning. The major contribution used Galton's height, with a complex application of algorithms with significantly high accuracy [5]. While on a different path, hybrid models introduced the inclusion of factors from different perspectives for overall improved predictive performance [3]. Contrarily, bone age estimation is shown to play a key role in height prediction. It shows its important applicability in all contexts [1].

Research also points out that the anthropometric measures of the body, such as leg length and sitting height, have been found to be the most important predictors to estimate adult

979-8-3315-2169-1/25/$31.00 ©2025 IEEE

stature [2]. Another piecewise linear regression approach has also been regarded as promising, and it can potentially create further research venues in determining model performance [6]. Further studies on growth patterns have thus confirmed the validity of parental stature as the best predictor for both genetic and environmental contributions towards the trajectory of the child's growth [4]. In addition to this, multivariate analyses have made various assumptions of linear models in stature estimation and formed a contribution basis for the understanding of height prediction analysis [7]. Therefore, through these studies, consecutively, it has highlighted the importance of merging the conventional approach with the new ideas to improve adult height predictions with high reliability and thereby finally would provide essential benefits for pediatricians and parents in monitoring children's growth.

III. Exploratory Data Analysis and Feature Extraction

A. Data Collection

Before conducting the experiments, we familiarized ourselves with the dataset. We chose Galton's height dataset, which contains records of 898 individuals from 19th-century Britain [5]. However, data like Gothenburg and Edinburgh data from 1974 and 1990 have longitudinal records, but this is also time-consuming to acquire. Galton's data, while determining the causes of growth, give both children's and parents' height with them simultaneously.Galton's height information comprises of six highlights:

- Father: Paternal height measured in inches.
- Mother: Mother's height measured in inches.
- Child: Child's height measured in inches.
- Gender: Indicates if the child is male or female.

B. Feature Engineering

This project aims to improve the accuracy of adult height prediction models by incorporating additional features through comprehensive feature engineering. The study goes beyond genetic factors to include socioeconomic, environmental, and lifestyle elements, aims to provide a deeper understanding of height variations and improve prediction methods.

- avg parent height: Aggregated parental stature used to improve height prediction algorithms.
- Birth order: The impact of birth order on growth patterns.
- Household income: Socioeconomic status as a factor influencing nutritional access and health care.
- living environment: An assessment of the influence of environmental factors on growth outcomes.
- Diet quality: An evaluation of dietary intake quality and its impact on physical development.
- Grandparent height: The incorporation of familial genetic and environmental contributions over generations.
- play sports: Measurement of physical activity levels and their relationship to growth trajectories.

C. Outlier Removal

As demonstrated in Table 1, Galton had taken data from 898 people originally. Outliers are excluded from this cleaned dataset that has 881 people as presented in Table 2. Assuming that cases may seem anomalous to a particular result, such as short parents having very tall children.

TABLE I
GALTON'S HEIGHT DATASET PRIOR TO OUTLIER REMOVAL

Statistic	*Father*	*Mother*	*Son*	*Daughter*
Count	898	898	898	898
Mean	69.2328	64.0844	0.5178	0.4822
Standard Deviation	2.4702	2.3070	0.4999	0.4999
Max	78.5000	70.5000	1.0000	1.0000
Q3 (75%)	71.0000	65.5000	1.0000	1.0000
Q2 (50%)	69.0000	64.0000	1.0000	0.0000
Q1 (25%)	68.0000	63.0000	0.0000	0.0000
Min	62.0000	58.0000	0.0000	0.0000

TABLE II
GALTON'S HEIGHT DATASET FOLLOWING OUTLIER REMOVAL

Statistic	*Father*	*Mother*	*Son*	*Daughter*
Count	889	889	889	889
Mean	69.2385	64.0751	0.5163	0.4836
Standard Deviation	2.4816	2.3168	0.5000	0.5000
Max	78.5000	70.5000	1.0000	1.0000
Q3 (75%)	71.0000	65.5000	1.0000	1.0000
Q2 (50%)	69.0000	64.0000	1.0000	0.0000
Q1 (25%)	68.0000	63.0000	0.0000	0.0000
Min	62.0000	58.0000	0.0000	0.0000

D. Distribution of Child's Height in Correlation with Parental Heights

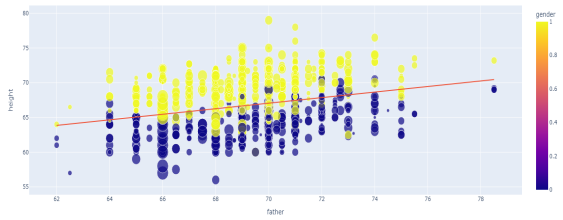

Fig. 1. Relationship Between the children's height and father's height Distribution

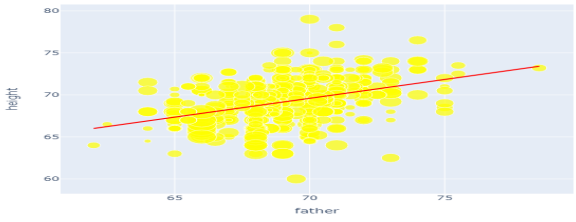

Fig. 2. Relationship Between the daughter's height and father's height Distribution

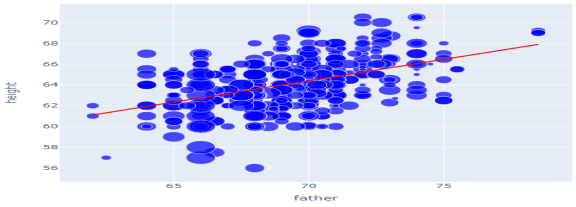

Fig. 3. Relationship Between the son's height and father's height Distribution

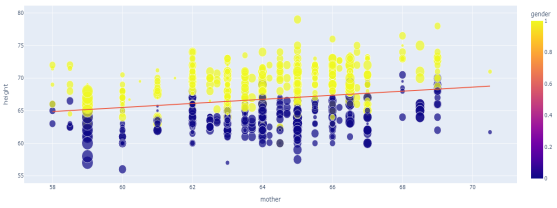

Fig. 4. Relationship Between the children's height and mother's height Distribution

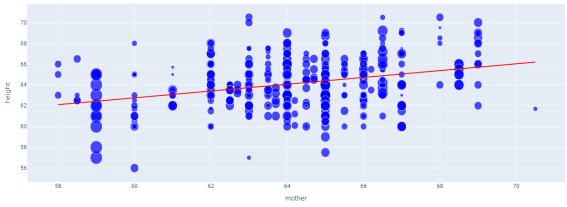

Fig. 5. Relationship Between the daughter's height and mother's height Distribution

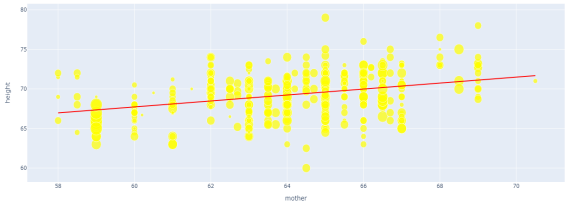

Fig. 6. Relationship Between the Son's Height and Mother's Height Distribution

The graph represents the heights of children in terms of their parents' using Galton's data. Figures 1 and 4 describe how both parents affect the height of children, while figures 2 and 5 show that girls are generally shorter than boys, thus portraying an even closer correlation between the height of a daughter with that of her father. Fig. 3 and 6 showed less correlation of mother's heights with her children's heights than father's heights, for sure due to some other kinds of genetic and environmental effects.

E. Correlation Analysis

A correlation approach is used to examine the height data from Galton. Mother and father heights, as well as MPH, are the parameters that are measured in the correlation analysis. The association coefficients between these variables and height are calculated individually for each child's gender.

For boys (cm):

$$\text{MPH} = \frac{\text{Father's Height} + \text{Mother's Height} + 13}{2}$$

For girls (cm):

$$\text{MPH} = \frac{\text{Father's Height} - 13 + \text{Mother's Height}}{2}$$

The correlation analysis was performed using Pearson's method. The formula for Pearson's correlation coefficient is:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where covariance is calculated as:

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Substituting, the correlation coefficient can be expressed as:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

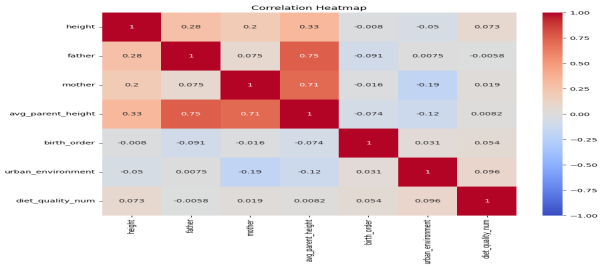

Fig. 7. Correlation Anaysis of child's height in Galton's Dataset (before removing outliers)

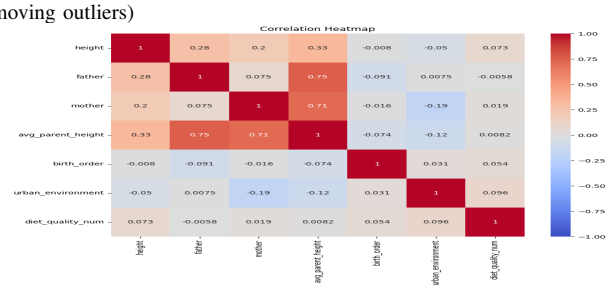

Fig. 8. Correlation Analysis of son's height in Galton's Dataset (before removing outliers)

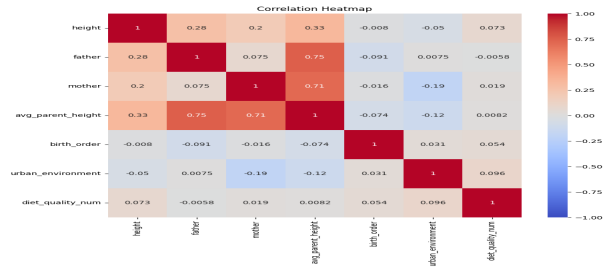

Fig. 9. Correlation Analysis of daughter's height in Galton's Dataset (before removing outliers)

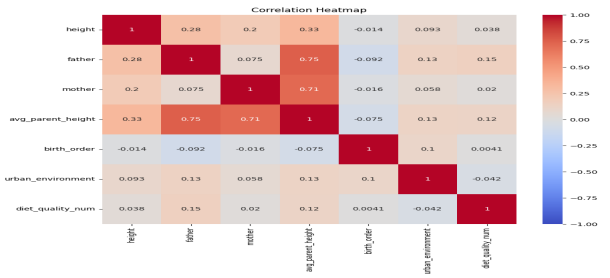

Fig. 10. Correlation Analysis of child's height in Galton's Dataset (after removing outliers)

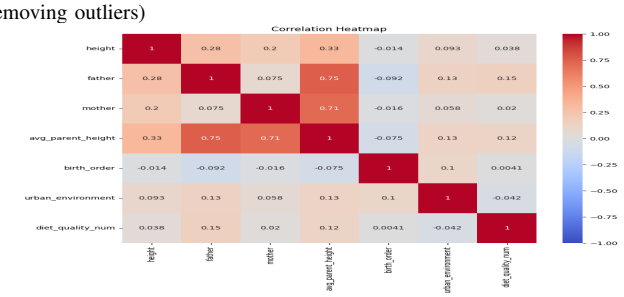

Fig. 11. Correlation Analysis of son's height in Galton's Dataset (after removing outliers)

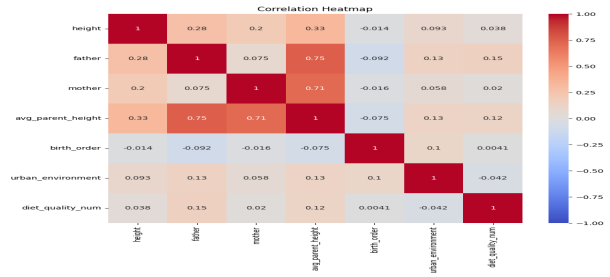

Fig. 12. Correlation Analysis of Daughter's Height in Galton's Dataset (after removing outliers)

From the data in height of Galton, one can infer that the height of a child owes much more to the father than to the mother. Analysis of Figures 7, 8, and 9 before elimination of outliers revealed that daughters are significantly affected by the height of the father more than sons. Figures 10, 11 and 12 confirm that the removal of outliers clears things up. Figure 11 shows that a son's height is significantly affected by the father but hardly at all by the mother. The difference between the influence of the father and the influence of the mother on the height of their respective sons in comparison to daughters is negligible in that both parents have a similar influence on the son's height once the outliers are removed.

Figure 12 illustrates that impacts from fathers to their daughters' height are stronger than those resulting from mothers, regardless of whether the child is male or female. The correlation analyses for both the pre and post-outlier data clearly state a direct relationship between the heights of parents and children. Parent-daughter height relations appear to be more linear as regards each other, meaning AHP could predict girls' heights better than boys'.

IV. Experiments And Results

A. Experiment 1

We first examine the Galton height dataset without any outliers removed, since their removal can impact the performance of any predictive model that is developed based on this particular dataset. We wish to lay a baseline understanding for the relationship between parents' heights and children's in the non-clean data. This includes the assessment of how well classical regression models can handle those relationships given no pre-processing of the data. We also explore the effect of outliers that may introduce distortion: highlighting variability in the data. The objective is to compute the unadjusted effect of parental height on child height with no statistical correction.

B. Experiment 2

The outlier reduction step will be included in the second experiment. With this, upon removing outliers from the Galton data set, then removes the outliers and re-assesses the relationships between the heights of the parents and their children, now under normal conditions. Outliers distort the accuracy of a model; therefore, removing the same should yield a cleaner and more reliable model. The cleansed data set will thus give a clearer view as to whether the parent has an effect on the child's height. This experiment will be compared with those results from Experiment 1 to analyze how the removal of outliers can impact machine learning models' performance.

C. Results of Experiment 1

The experiment was conducted for the possibility of the forecast of child height from Galton's height dataset without removing outliers, by comparing learning-based models - namely, Linear Regression, SVR, XGBoost, and LightGBM - with the traditional MPH method. The use of advanced techniques and creation of new feature involved factors like average parental height, birth order, and sports participation.

The result of all machine learning models surpassed the MPH method, which is a proof of the existence of additional variables. The addition of extreme parental heights introduced noise, which led to a degradation in performance for the models; yet still, the models were sound, particularly when feature engineering was applied. Experiment 1 fared better in the results with all different machine learning models prior to applying any method of outlier removal, whereas more analysis took place in Experiment 2.

TABLE III
RSME on Original Data and Preprocessed Data (Before Removing Outliers)

Model	Original Data	Before Outlier Removal
Linear Regression	3.47	2.33
SVR (linear)	3.48	3.22
XGBoost	3.74	3.26
LightGBM	3.56	3.08
MPH	3.47	3.20

TABLE IV
RSME ON ORIGINAL DATA AND PREPROCESSED DATA (AFTER REMOVING OUTLIERS)

Model	Original Data	After Outlier Removal
Linear Regression	3.47	2.20
SVR (linear)	3.48	2.22
XGBoost	3.74	2.20
LightGBM	3.56	2.03
MPH	3.47	2.20

D. Results of Experiment 2

Building on the insights from Experiment 1, Experiment 2 evaluates regression models using Galton's height data after outlier removal. The results suggest better predictability for all models. These include Linear Regression, SVR, XGBoost, MPH, and LightGBM. Although SVR and XGBoost depend on data distribution and thus marginally outperform other algorithms, stronger is a machine learning algorithm. Confirming the hypothesis put forward by Experiment 2, outlier removal brings significant improvements in the accuracy of the regression model as compared with baseline methods.

E. Model performance before AND after outlier removal

TABLE V
MODEL PERFORMANCE BEFORE AND AFTER OUTLIER REMOVAL

Model	Before Outlier Removal	After Outlier Removal
Linear Regression	60.6	63.3
SVR (linear)	59.7	62.7
XGBoost	62.7	63.2
LightGBM	66.2	68.8
MPH	60.6	63.3

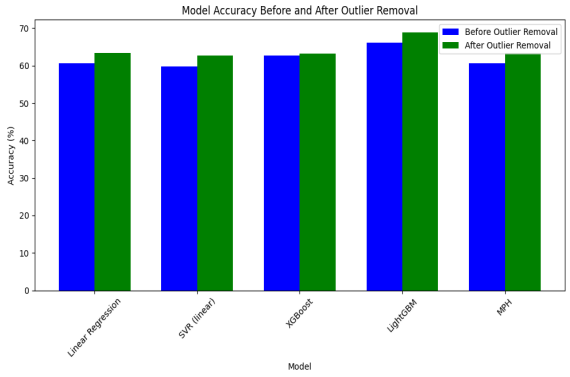

Fig. 13. Model Accuracy Before and After Outlier Removal

Fig 13 Outlier detection and removal accuracies of classifiers on Galton dataset Experiment Outlier removal Experiment to rate Model Before After Model performance, The respective accuracies of classifiers on Galton dataset before and after outlier removal are shown in Fig 13. Experiment 1 experimented the classifiers on the original dataset whereas Experiment 2 again experimented the classifiers after removing the outliers from it.

V. COMPARATIVE ANALYSIS

TABLE VI
TRAINABLE AND TESTABLE PARAMETERS FOR EACH MODEL

Model	Trainable Parameters	Testable Parameters
Linear Regression	10	178
SVR (linear)	10	178
XGBoost	700	178
LightGBM	3200	178
MPH Model	10	178

TABLE VII
TRAINING AND TESTING ACCURACY FOR VARIOUS MODELS

Model	Training Accuracy	Testing Accuracy
Linear Regression	0.6452	0.6409
SVR (linear)	0.6417	0.6363
XGBoost	0.9913	0.6463
LightGBM	0.8906	0.6920
MPH Model	0.6452	0.6409

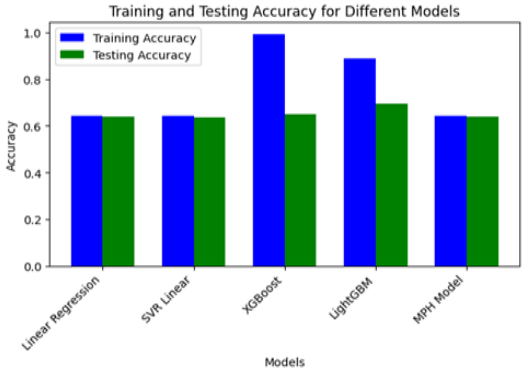

Fig. 14. Training and Testing Accuracy for Different Models

TABLE VIII
PRECISION, RECALL, F1-SCORE, AND ACCURACY FOR VARIOUS MODELS

Model	Precision	Recall	F1-Score	Accuracy
Linear Regression	0.8314	0.9024	0.8654	0.6409
SVR (linear)	0.8202	0.8902	0.8538	0.6363
XGBoost	0.7912	0.8780	0.8323	0.6462
LightGBM	0.8352	0.8658	0.8502	0.6920
MPH Model	0.8314	0.9024	0.8655	0.6409

VI. CONCLUSION

This research discusses the advantages of applying machine learning models to predict child height, as demonstrated by an evaluation of Galton height data. Experiment 1 showed that integrating extra factors including average parental height, birth order, and family wealth enhanced model accuracy when compared to the traditional Mid-Parental Height (MPH) technique. These variables allowed the models to better represent the complex interaction of genetic and environmental factors that affect height.

Experiment 2 highlighted the value of data preprocessing by showing that removing outliers improved prediction accuracy. This underscores the importance of data cleaning for

enhancing model performance. Overall, the study underscores the benefits of advanced machine learning and comprehensive feature engineering in achieving more accurate predictions, providing practical insights for developing robust predictive models in similar contexts.

VII. Future Work

Future research will focus on incorporating temporal variables to account for dynamic changes throughout an individual's growth period. Currently, factors like diet, physical activity, and healthcare interventions are treated as static, but they vary over time and significantly influence growth. Longitudinal data that tracks these factors at different stages of growth could provide a more accurate representation of an individual's development. Employing time-series analysis will enable the model to adapt to these fluctuations, capturing their impact on height prediction over time. Additionally, dynamic feature engineering will be explored, where factors like diet quality, exercise levels, and medical treatment evolve throughout a child's growth phases. Real-time data from wearable devices and health tracking apps could be integrated to provide personalized and up-to-date insights. Incorporating environmental and societal changes, such as shifts in climate or economic conditions, will further refine the model for more accurate predictions.

References

[1] J. Suh, J. Heo, S. J. Kim, S. Park, M. K. Jung, H. S. Choi, Y. Choi, J. S. Oh, H. I. Lee, M. Lee, K. Song, A. Kwon, H. W. Chae, and H.-S. Kim, "Bone Age Estimation and Prediction of Final Adult Height Using Deep Learning," *Yonsei Medical*, vol. 64, no. 11, pp. 1150–1159, Nov. 2023.

[2] J.-H. Park, M. Lee, D. Kim, H.-W. Kwon, Y.-J. Choi, K.-R. Park, S. Park, S.-B. Park, and J. Cho, "Estimating Adult Stature Using Metatarsal Length in the Korean Population: A Cadaveric Study," *Int. J. Environ. Res. Public Health*, vol. 19, no. 22, p. 15124, Nov. 2022.

[3] K. Mao, L. Chen, X. Fan, J. Mao, X. Zhou, and K. Fang, "Lsalo-Bp: A Hybrid Model for Children Adulthood Height Prediction," *Zhejiang University of Technology*, Jan. 2022.

[4] A. Holmgren, A. Niklasson, A. F. M. Nierop, G. Butler, and K. Albertsson-Wikland, "Growth Pattern Evaluation of the Edinburgh and Gothenburg Cohorts by QEPS Height Model," *Pediatric Res.*, vol. 92, no. 2, pp. 592–601, Aug. 2022.

[5] M. Shmoish, A. German, N. Devir, A. Hecht, G. Butler, A. Niklasson, K. Albertsson-Wikland, and Z. Hochberg, "Prediction of Adult Height by Machine Learning Technique," *Clinical Endocrinology & Metabolism*, vol. 106, no. 7, pp. 301–310, Jan. 2021.

[6] A. Bemporad, "A Piecewise Linear Regression and Classification Algorithm with Application to Learning and Model Predictive Control of Hybrid Systems," *IEEE Trans. Autom. Control*, vol. 68, no. 6, pp. 3194–3209, Jun. 2023.

[7] H. B. Khazri, S. C. Shimmi, and M. T. H. Parash, "A Multivariate Analysis to Propose Linear Models for the Stature Estimation in the Sabahan Young Adult Population," *PLoS ONE*, vol. 17, no. 8, Art. no. e0273840, Aug. 2022.

[8] M. Maes, M. Vandeweghe, M. Du Caju, Ch. Ernould, J.-P. Bourguignon, and G. Massa, "A Valuable Improvement of Adult Height Prediction Methods in Short Normal Children," *Hormone Research*, vol. 48, no. 8, pp. 555–561, Jan. 1997.

[9] K. Umapavankumar, S. V. N. Srinivasu, S. Nageswara Rao, and S. N. Thirumala Rao, "Machine learning usage in Facebook, Twitter and Google along with the other tools," in *Emerging Research in Data Engineering Systems and Computer Communications: Proceedings of CCODE 2019*, pp. 465–471. Singapore: Springer Singapore, 2020.

[10] S. S. N. Rao, Y. S. Krishna, and K. N. Rao, "A survey: routing protocols for wireless mesh networks," *International Journal of Research and Reviews in Wireless Sensor Networks (IJRRWSN)*, vol. 1, no. 3, pp. 23–29, 2011.

[11] S. S. N. Rao, Y. S. Krishna, and K. N. Rao, "Performance Evaluation of routing protocols in Wireless Mesh networks," *International Journal of Computer Applications*, vol. 68, no. 7, pp. 12–18, 2013.

[12] G. Parimala, S. Nageswararao, and K. LakshmiNadh, "DDSRC: Algorithm for improving QOS in VANET," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 7, pp. 1327–1331, 2019.

[13] S. S. N. Rao, Y. S. Krishna, and K. N. Rao, "An elliptical routing protocol for wireless mesh networks: Performance analysis," *International Journal of Computer Applications*, vol. 102, no. 8, pp. 24–31, 2014.

[14] S. S. N. Rao, Y. S. Krishna, and K. N. Rao, "Notice of Removal: Active topology based routing approaches for Wireless Mesh Networks," in *2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)*, pp. 1–3. IEEE, Jan. 2015.

[15] S. S. N. Rao, Y. S. Krishna, and K. N. Rao, "A study of routing metrics for wireless mesh networks," *International Journal of Research and Reviews in Wireless Communications (IJRRWC)*, vol. 1, no. 2, pp. 19–25, 2011.