AI-Driven ADHD Prediction and Analysis at Early Age:

A Novel Approach Integrating Machine Learning, Explainable AI, LLMs, and Dialogflow with a Virtual Therapist Chatbot (ComfortChat)

Methodology

Senior Design Project

Md. Adham Wahid	ID# 2111177042
Fatema Afsan Ema	ID# 2022281642
Istiak Ahasan	ID# 2012082642

Faculty Advisor:
Dr. Sifat Momen
Associate Professor
ECE Department

Department of Electrical and Computer Engineering
North South University

Spring 2025

Methodology of the Proposed System

0.1 Methodology

This section presents a comprehensive description of the proposed framework for detecting and analyzing Attention Deficit Hyperactivity Disorder (ADHD), as outlined in Figure 1. It encompasses the key components of the system, including data collection, preprocessing, machine learning-based classification, large language model (LLM) integration, explainable AI (XAI) techniques, and system deployment through a Dialogflow-powered therapeutic chatbot.

The methodology follows a logical and sequential flow, beginning with data acquisition and concluding with the deployment of the complete system into both web and mobile platforms. Several components are interconnected and function as intermediary bridges between different modules, ensuring a cohesive and modular system design.

This structured approach not only defines the backbone of the research but also establishes the essential mechanisms that drive each phase of the system. The proposed methodology is novel in its integration of both predictive analytics and real-time interactive support, making it, to the best of our knowledge, the first of its kind to offer a unified solution that combines machine learning predictions with an intelligent, chatbot-based therapeutic interface.

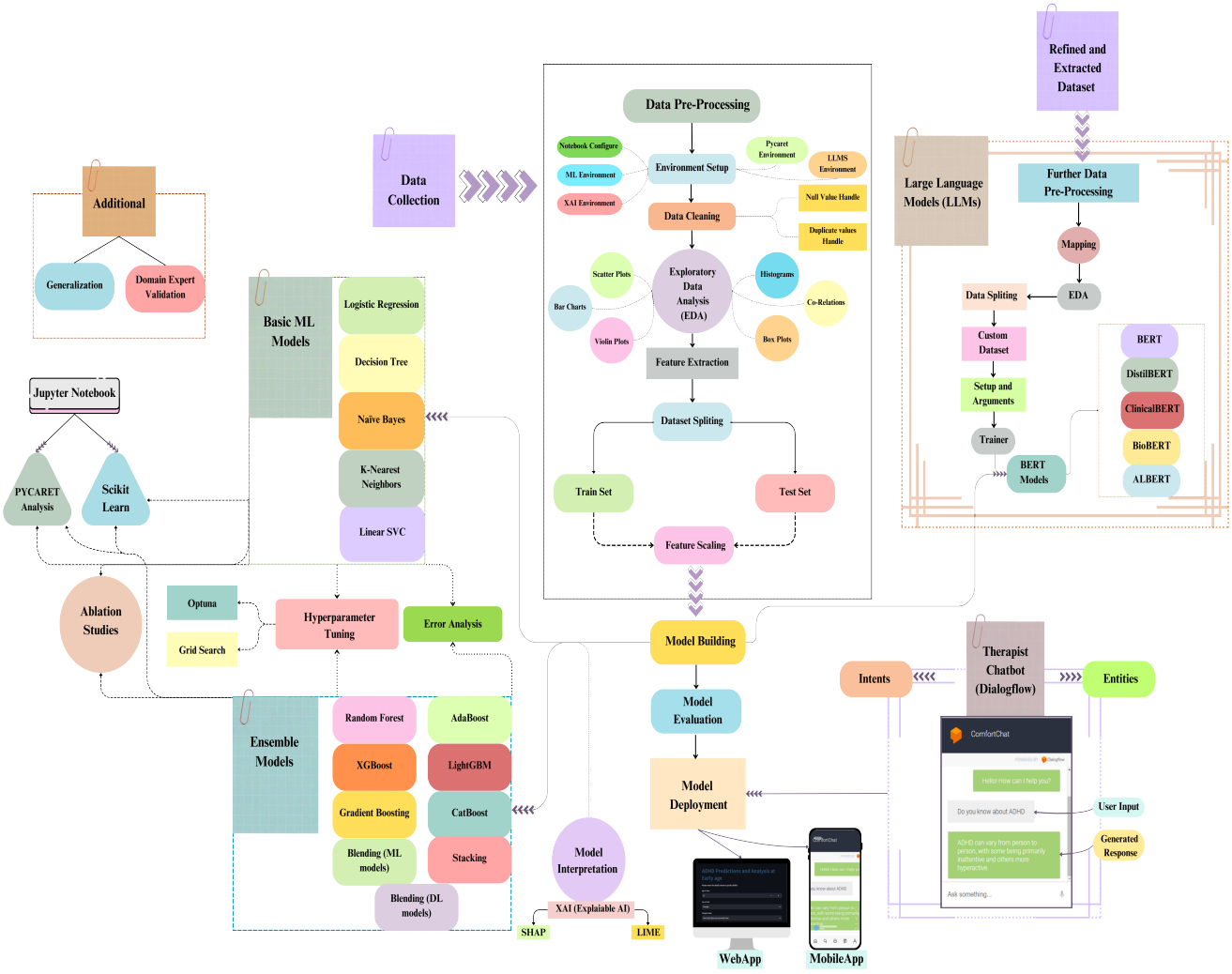

Figure 1: Methodology diagram of the proposed system.

0.1.1 Data collection

Data collection was conducted through a formal request to the Child and Adolescent Health Measurement Initiative (CAHMI) [1], housed at Johns Hopkins Bloomberg School of Public Health [2]. CAHMI is a leading authority in child health data, supporting initiatives such as ACEs research, the Maternal and Child Health Measurement Network, and the National Survey of Children's Health (NSCH).

For this study, we primarily used the 2021–2022 NSCH dataset, comprising 104,995 rows and 826 variables. After filtering based on literature and SPSS codebook review, a refined dataset of 73 relevant features was used for model training. To evaluate generalizability, we also employed two additional datasets: the 2022–2023 NSCH (most recent) and the 2018–2019 NSCH (historical reference).

0.1.2 Data pre-processing

Data preparation includes data preprocessing, which is any kind of processing done on raw data to prepare it for another data processing step. This study has applied and explored different types of data preprocessing techniques. They include environment setup, data cleaning, exploratory data analysis, feature extraction, dataset splitting, and feature scaling.

Environment setup

To keep track and ensure smooth transition environment setup is a must. Have used the Jupyter Notebook as the coding platform . Also set up five different environments includes the ML environment, XAI environment , Pycaret environment, LLMs environment and general configured environment. The Figure 2 shows the used environments for this study.

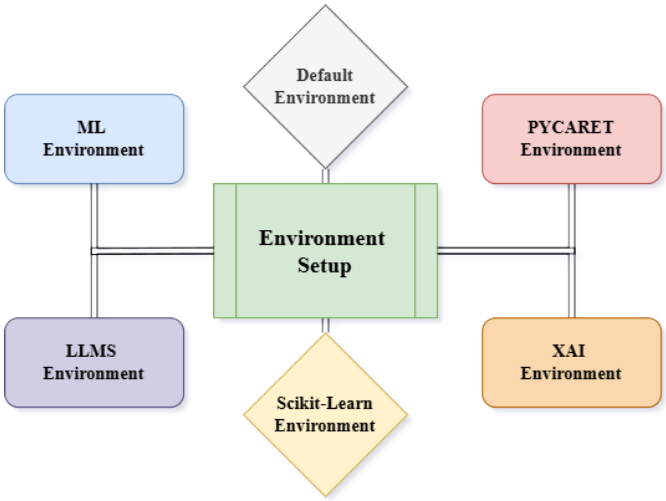

Figure 2: Environment setup

Data cleaning

The dataset comes with a code-book where a detailed information about the survey has been illustrated . Initially have analyzed the code-book and found out there are some specific encoded values which defines the missing values . To be more specific the values are [90,95,96,99]. As the dataset has decent amount of values , so decided to apply both null value extraction and imputation through the most frequent value . Extracted [95,96,99] holding values and imputed the values holding the key value 90. The dataset size turned into 67336 rows x 73 columns containing all the relevant features. Duplicate values has been also plucked out for ensuring unbiased result . The dataset size been updated to 67171 rows x 72 columns removing another redundant feature . Proper

cleaning of the data [3]guarantees that data is in its appropriate form depicting clearly the current patterns and relationships. Better reliability, there is an enhancement of data quality towards achieving that it is more objective and accurate. Aside from that, this attempt to clean enhances the strength of analytics from the data, we now ensure that subsequently constructed models were based on trust worthy data. This prepared data on an additional cleansing with our analysis can move into a good. a more advanced level of analysis and create good predictive models for the sake of deriving useful and relevant results. The Figure 3 illustrates the data cleaning process in the data pre-processing.

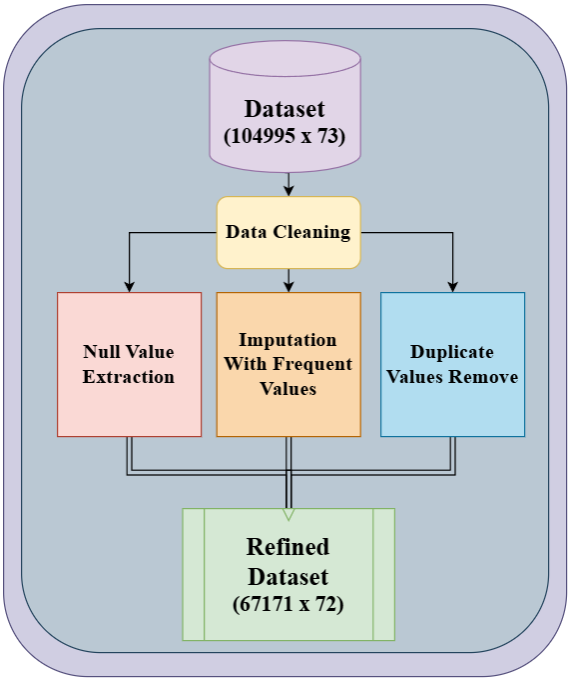

Figure 3: Data cleaning process

Encoding attributes

Encoding [4] is a crucial part during the data pre-processing . Fortunately, received the encoded dataset version from CAHMI [1] . The dataset has a code-book that illustrates the feature informations and relevant data.

Exploratory data analysis (EDA)

Demograpchic and Socioeconomic analysis

While analyzing and extracting the key characteristics of ADHD, the socioeconomic and demographic features play a crucial role. Simply it's a key element for the insights and proper understandings. Demographic features like age of the child, age of the mother, the

family structure of the child, race and ethnicity, financial situation of the family, number of family members in the family, and parental nativity are very basic and solid attributes for the insights.

Table 1: Socio-Demographic Distribution (AGE and GENDER)

Category	**Group**	**Frequency (%)**
SC_AGE_YEARS (Age of child)	3-5	17,870 (26.81%)
	6-12	26,259 (39.39%)
	13-17	22,537 (33.81%)
	Total	66,666 (100.00%)
MOMAGE (Age of mother)	18-25	13,890 (20.84%)
	26-35	38,105 (57.18%)
	36-45	14,671 (22.01%)
	Total	66,666 (100.00%)
sex_2122 (Gender Distribution)	Male	34,530 (51.80%)
	Female	32,136 (48.20%)
	Total	66,666 (100.00%)

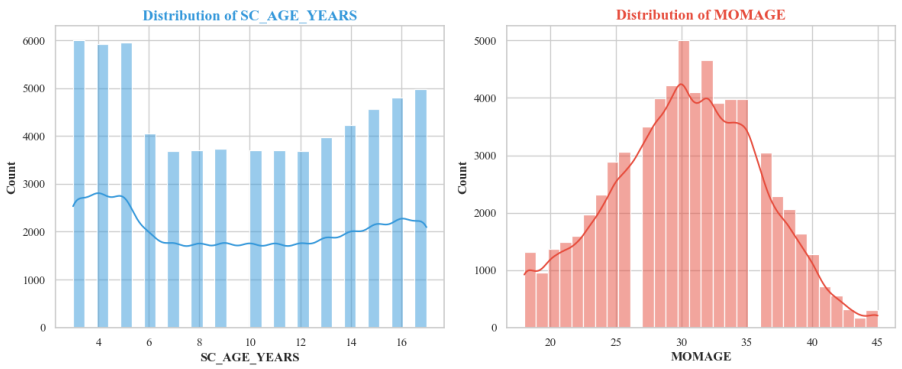

Figure 4: Age distribution of Child and Mother

From the Table 1 and Figure 4, the intuition of the ages for both the children and mothers can be retrieved. There are balanced scatter of the child age distribution, while the mother's age seem to have density in the 26-35 years age group reflecting typical childbearing age patterns. The higher proportion falls within the school-going age in the children group section. While still considerable, the percentage of older moms and young moms are comparatively smaller.

Table 2: Race and family structure distribution with ADHD positive cases

Category	Group	Frequency (%)
race4_2122 (Race)	White	44,952 (67.41%)
	Hispanic	9,071 (13.61%)
	Other	9,036 (13.56%)
	Black	3,607 (5.42%)
	Total	66,666 (100.00%)
ADHD Positive (by Race)	White	5,211 (11.59%)
	Hispanic	842 (9.28%)
	Other	729 (8.07%)
	Black	393 (10.90%)
famstruct5_2122 (Family Structure)	Two Parents (Married)	48,321 (72.50%)
	Single Parent	12,888 (19.34%)
	Two Parents (Unmarried)	3,480 (5.22%)
	Grandparent	1,486 (2.23%)
	Other	491 (0.71%)
	Total	66,666 (100.00%)
ADHD Positive (by Family Structure)	Two Parents (Married)	4,498 (9.31%)
	Single Parent	1,805 (14.00%)
	Two Parents (Unmarried)	413 (11.87%)
	Grandparent	247 (16.63%)
	Other	49 (9.98%)

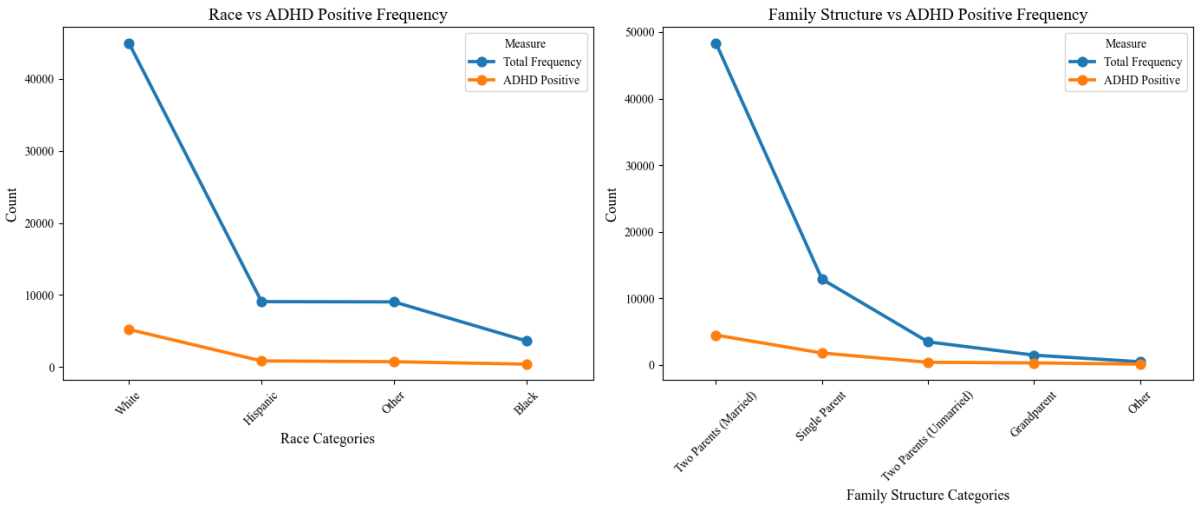

Figure 5: Race x Family structure distribution (ADHD)

The analysis reveals that ADHD-positive cases are most prevalent among White children and those from two-parent married households when considering absolute frequencies. However, Figure 5 demonstrates that prevalence rates tell a more nuanced story: while the absolute counts are lower, children in grandparent-headed households show the highest ADHD rate (16.63%), with minimal variation between this and other family structures. Complete demographic breakdowns by race and family structure are provided in Table 2.

Adverse Childhood Experiences (ACEs) and Social Support

Table 3: Overall and ADHD positive frequency adverse childhood experiences and social support

Feature	Overall	ADHD Positive	Percentage (%)
ACE2more6HH_2122 (Household Based)			
No Household ACE	45,977	3,489	7.59%
1 Household ACE	12,083	1,688	13.97%
2+ Household ACEs	8,606	1,998	23.22%
ACE1more4Com_2122 (Community Based)			
No Community ACE	59,954	5,277	8.80%
1+ Community ACEs	6,712	1,898	28.28%
ACEincome_2122 (Financial Hardship)			
Never	40,634	3,352	8.25%
Rarely	19,222	2,488	12.95%
Often	5,590	1,040	18.60%
Very Often	1,220	295	24.18%
ACESexDiscrim_2122 (Discrimination Based)			
No	65,597	6,854	10.45%
Yes	1,069	321	30.03%
ACEdivorce_2122 (Parental Separation)			
No	52,041	4,549	8.74%
Yes	14,625	2,626	17.95%
ACEdrug_2122 (Substance Abuse in Household)			
No	60,305	5,779	9.58%
Yes	6,361	1,396	21.95%
EmSFamily_2122 (Emotional Support from Family)			
Family	48,687	5,197	10.67%
Not Family	5,775	544	9.42%
EmSupport_2122 (General Emotional Support)			
Yes	54,462	5,741	10.55%
No	12,204	1,434	11.75%

Table 3 and Figure 6 signifies the overall and ADHD positive frequency and their distribution based on Adverse childhood experiences and social support factors. Considering the factors it can be stated that gender discrimination, parental separation, drugs, lack of family and emotional support have an impact on ADHD cases. ADHD rate jumps from 8.80% (none) to 28.28% (1+ ACEs), highlighting community adversity's impact. ADHD-positive rate is 30.03% for those experiencing discrimination, the highest among all factors.

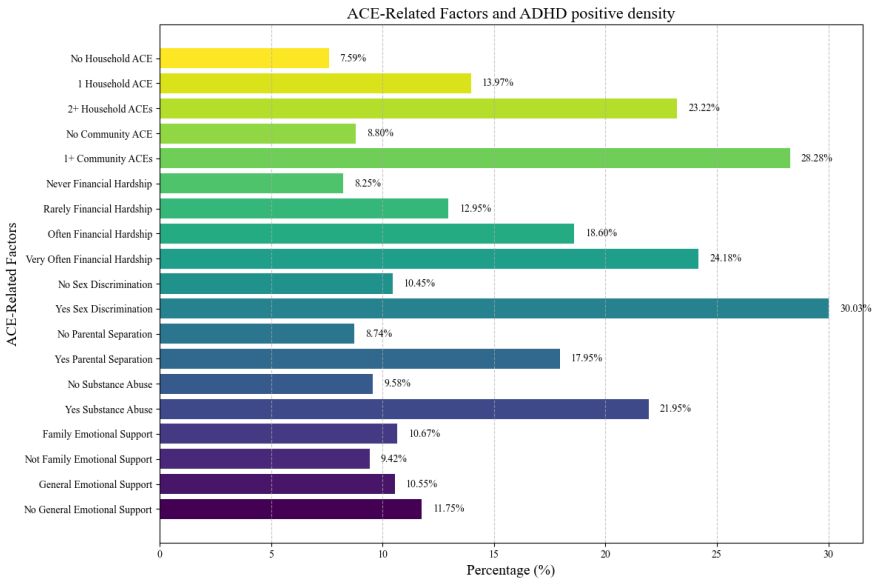

Figure 6: ACEs and ADHD positive density

Health and medical features

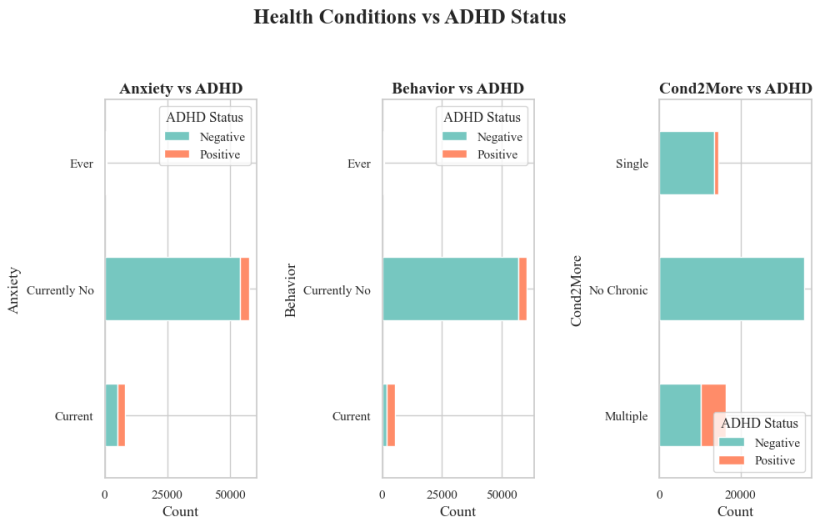

Figure 7: Top health contributors for ADHD

According to this study and survey, (anxiety, behavior, and chronic diseases) are the top health-contributing features for ADHD-positive cases. Figure 7 shows ADHD-free and ADHD-bearing ratios for those features. Having multiple health issues push towards ADHD. Severe behavior issues and ever had cases are also responsible for ADHD cases which is also the same for anxiety considering the Figure 7.

Environmental and lifestyle analysis

Table 4: Overall and ADHD Positive Frequency for Selected Lifestyle and Environment Features

Combinations	Overall Count	ADHD Positive	Percentage (%)
Fruit Intake (fruit_2122)			
2 times/day	54,658	6,916	12.65%
3+ times/day	3,330	67	2.01%
1 time/day	3,310	52	1.57%
4-6 times/month	3,266	72	2.20%
1-3 times/month	1,779	51	2.87%
Never	323	17	5.26%
Vegetable Intake (vegetables_2122)			
2 times/day	53,080	6,880	12.96%
3+ times/day	1,181	36	3.05%
1 time/day	3,747	62	1.65%
4-6 times/month	3,550	59	1.66%
1-3 times/month	4,069	85	2.09%
Never	1,039	53	5.10%
Sugary Drink Intake (SugarDrink_2122)			
No	56,249	6,960	12.38%
1-3 times/month	6,929	123	1.77%
4-6 times/month	1,441	35	2.43%
1 time/day	1,198	22	1.83%
2 times/day	618	20	3.24%
3+ times/day	231	15	6.49%
Breastfeeding History (BrstEver_2122)			
Yes	63,844	7,064	11.07%
No	2,822	111	3.93%
Exclusive Breastfeeding (ExBrstFd_2122)			
6+ months	54,404	6,900	12.69%
4-6 months	5,349	73	1.36%
4 months or less	4,091	91	2.22%
Never	2,822	111	3.93%
Park Access (park_2122)			
Yes	51,108	5,300	10.37%
No	15,558	1,875	12.05%
Library Access (library_2122)			
Yes	44,780	4,699	10.50%
No	21,886	2,476	11.32%
Smoking Household (smoking_2122)			
No	58,582	5,874	10.03%
Yes	8,084	1,301	16.09%

This part of features are interesting. Exploring these features it has been seen that most of the children are grown with great care. Table 4 shows the frequency of overall and ADHD-positive cases based on lifestyle and environmental features. The ratio of normal and ADHD cases are kind of similar for lifestyle and environmental features.

Behavioral and developmental exploration

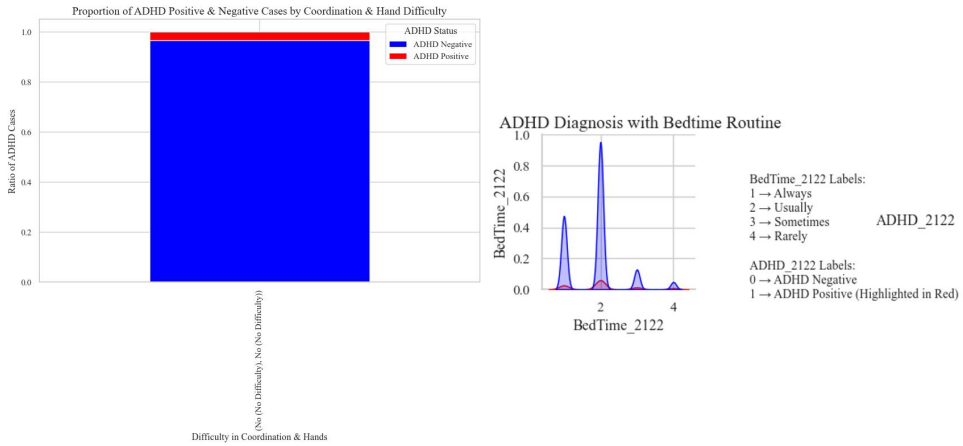

Figure 8: Bedroutine & Coordination impact on ADHD

Bed-routine, coordination and hand difficulty has impact on ADHD as shown in Figure 8. Most children follow a specific bed routine. Among them a portion has ADHD. But most child shows normal behavior. Among 23.90% children who don't follow bedtime get ADHD positive cases according to the survey. The coordination and hand difficulties have a mild impact on the ADHD.

0.1.3 Feature Selection Pipeline

Feature selection plays a critical role in improving model performance and reducing computational complexity, particularly when dealing with high-dimensional datasets. For this study, a multi-step feature selection strategy was employed to refine the initial dataset.

The process involved:

Correlation Matrix (Pearson) — to identify weakly correlated features with the target variable.

Variance Inflation Factor (VIF) — to detect and eliminate multicollinearity.

Recursive Feature Elimination (RFE) — to iteratively select the most relevant predictors based on model performance.

As illustrated in Figure 9, this pipeline ensured that only informative and non-redundant features were retained for model training. A total of 39 features with low predictive power or high multicollinearity were removed.

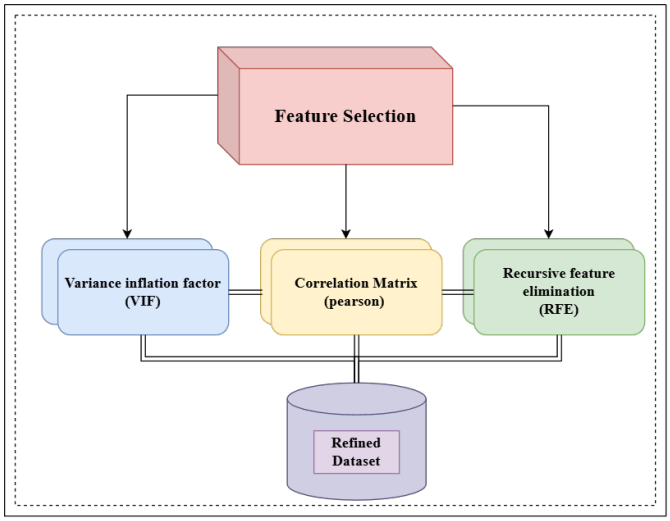

Figure 9: Feature selection workflow

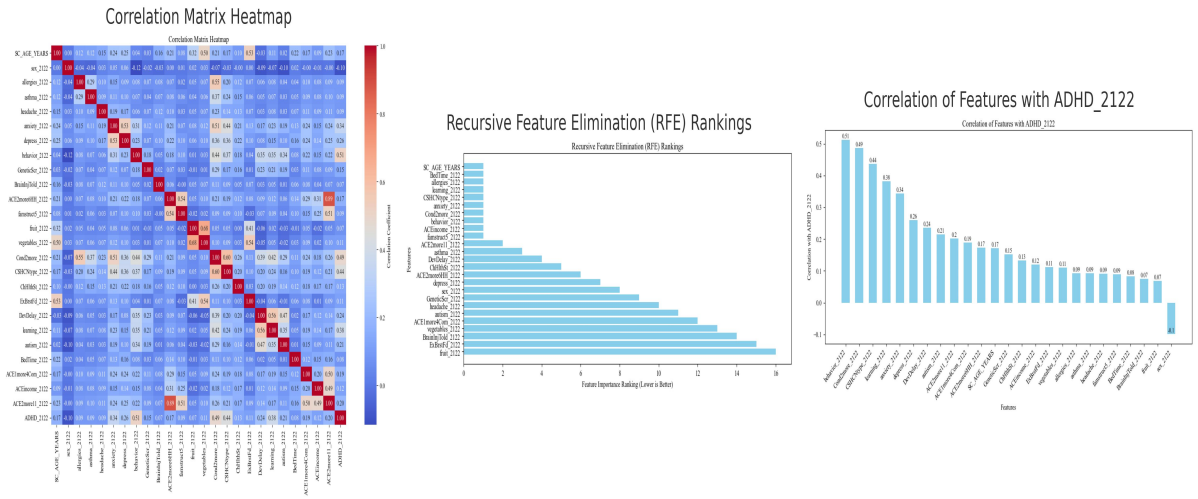

Figure 10: Selected Features Heatmap and Rankings

The refined dataset contains the mentioned features like the Table 5 and 26 features to be exact with 66,666 entries . The Figure 10 illustrates the correlation heatmap, recursive feature elimination rankings. Primarily, it can be stated that the behavior and chronic disease features put impact on the ADHD (Target) variable. For the RFE, the lower values are more impactful towards the target variable. VIF indicates the multicollinearity and the features that are more close to each other has been removed from the dataset for generating the precise and unbiased outcome.

Refined extracted dataset

Table 5: Final Refined Dataset

Feature Name	Relevant Question	Result Combination
SC_AGE_YEARS	What is the child's age?	Continuous
sex_2122	What is the child's gender?	Male, Female
allergies_2122	Has the child ever had allergies?	No, Ever, Current
asthma_2122	Has the child ever been diagnosed with asthma?	No, Ever, Current
headache_2122	Does the child experience headaches?	No, Ever, Current
anxiety_2122	Has the child ever been diagnosed with anxiety?	No, Ever, Current
depress_2122	Has the child ever been diagnosed with depression?	No, Ever, Current
behavior_2122	Does the child exhibit behavioral issues?	No, Ever, Current
GeneticScr_2122	Has the child undergone genetic screening?	Never, Ever, Identified
BrainInjTold_2122	Has the child been diagnosed with a brain injury?	Never, Ever, Had
ACE2more6HH_2122	How many adverse childhood experiences (household-based) has the child faced?	No, Single, Multiple
famstruct5_2122	What is the child's family structure?	2-parents, Single parent, Grandparent, Other
fruit_2122	How often does the child consume fruits?	No, 1-3x/m, 4-6x/m, 1x/day, 2x/day, 3+x/day
vegetables_2122	How often does the child consume vegetables?	No, 1-3x/m, 4-6x/m, 1x/day, 2x/day, 3+x/day
Cond2more_2122	How many health conditions does the child have?	None, Single, Multiple
CSHCNtype_2122	Does the child have special health care needs?	4 categories
ChHlthSt_2122	How is the child's overall health?	Good, Poor
ExBrstFd_2122	Was the child exclusively breastfed?	Never, Less than 6m, 6m, Greater than 6m
DevDelay_2122	Has the child been diagnosed with developmental delay?	No, Ever, Current
learning_2122	Does the child have a learning disability?	No, Ever, Current
autism_2122	Has the child been diagnosed with autism?	No, Ever, Current
BedTime_2122	Does the child have a consistent bedtime routine?	Always, Usually, Sometimes, Rarely
ACE1more4Com_2122	How many adverse childhood experiences (community-based) has the child faced?	No, Single, Multiple
ACEincome_2122	How often does the family struggle with income?	Never, Rarely, Somewhat, Very Often
ACE2more11_2122	Has the child experienced multiple adverse childhood experiences?	Single/None, Multiple
ADHD_2122 (Target)	Has the child been diagnosed with ADHD?	Negative, Positive

Total Frequency: 66,666

Table 5 picturize the portrait of the refined final dataset along with relevant questions, feature names and result combinations. The dataset size is 66,666 rows x 26 columns to be exact. It has been used for the ML,LLMs, XAI models training and evaluations. It holds a significant role in this study.

Data splitting

To prepare the data for model training and evaluation, the feature set was first separated from the target variable. This was accomplished by dropping the target column from the dataset to form the feature matrix, while the target variable was extracted from the corresponding column.

Subsequently, the dataset was split into training and testing subsets using the train_test_split function from the sklearn.model_selection module [5]. An 85:15 split ratio was applied, allocating 85% of the data for training and 15% for testing. To ensure reproducibility and consistent shuffling of the data, a fixed random seed of 42 was used.

This split strategy ensures an unbiased and balanced evaluation of the model's performance. The resulting training set comprised 56,666 rows $\times$ 25 features, while the test set included 10,000 rows $\times$ 25 features.

Feature scaling

To ensure that the features in the dataset are on a comparable scale and to improve the performance of the machine learning model, the Standard Scaler has been applied [6]. The Standard Scaler standardizes the features by transforming them such that they have a mean of 0 and a standard deviation of 1. This process ensures that each feature contributes equally to the model, preventing features with larger numerical ranges from disproportionately influencing the model's behavior. Standardization is particularly important for models that rely on distance metrics, such as linear regression, logistic regression, and support vector machines, as it ensures that no single feature dominates due to its scale. The given equation represents the standard scaler.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

0.1.4 Applied Machine Learning Algorithms

Fourteen machine learning models, including both basic and ensemble methods, were implemented to classify ADHD cases. Key mathematical formulations are included to support theoretical clarity.

Logistic Regression

A linear classification model using the sigmoid function to estimate class probabilities [7].

$$P(Y = 1|X) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{2}$$

Decision Tree

A rule-based learner that splits data using Gini impurity or Entropy to maximize information gain [8].

$$Gini = 1 - \sum_{i=1}^{n} p_i^2, \quad Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3}$$

K-Nearest Neighbors (KNN)

A non-parametric algorithm that classifies based on majority vote among the k closest points [9].

$$y = \arg\max_{c \in C} \sum_{i=1}^{k} \mathbb{1}(y_i = c) \tag{4}$$

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2} \tag{5}$$

Naïve Bayes

A probabilistic classifier based on Bayes' Theorem, assuming conditional independence between features [10].

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{6}$$

Linear SVC

A linear classifier that finds the hyperplane maximizing margin between classes [11]. Often used for binary classification and outlier detection.

Random Forest

An ensemble method that combines multiple decision trees trained on different data subsets to improve accuracy [12].

Gradient Boosting

A sequential technique where each learner tries to reduce the residual error of its predecessor [13].

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{7}$$

Extreme Gradient Boosting (XGBoost)

A scalable, regularized version of gradient boosting optimized for performance and speed [14, 15].

$$L_{xgb} = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{8}$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda |w|^2 \tag{9}$$

AdaBoost

Combines weak classifiers by assigning higher weights to misclassified instances [16].

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{10}$$

LightGBM

An optimized gradient boosting method using histogram-based learning and leaf-wise tree growth [17].

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} w_j h_j(x) \tag{11}$$

CatBoost

A gradient boosting algorithm developed for categorical feature handling, using ordered boosting to prevent overfitting [18].

$$G_t(x) = G_{t-1}(x) + \alpha \sum_{k=1}^{K_t} v_k g_k(x) \tag{12}$$

Stacking

An ensemble method that combines predictions of multiple base models using a meta-learner [19]. Improves accuracy by learning how to best weight each model's output.

Blending (ML)

Similar to stacking, blending uses a holdout validation set to train a meta-model, often using logistic regression [20].

Blending (DL)

An ensemble of deep learning architectures (e.g., CNN, RNN, Transformer) combined via a meta-model. Helps generalize predictions while minimizing overfitting [21].

0.1.5 Hyperparameter Optimization

Hyperparameter optimization is required to improve the performance of machine learning and transformer models, especially in applications such as ADHD prediction, where reliability and accuracy are necessary. In this study, we used both GridSearchCV and Optuna for optimizing the hyperparameters of various machine learning algorithms. These methods systematically look for the best hyperparameter configurations in order to enhance model performance. Figure 11 illustrates the used hyperparameter tuning techniques for machine learning models.

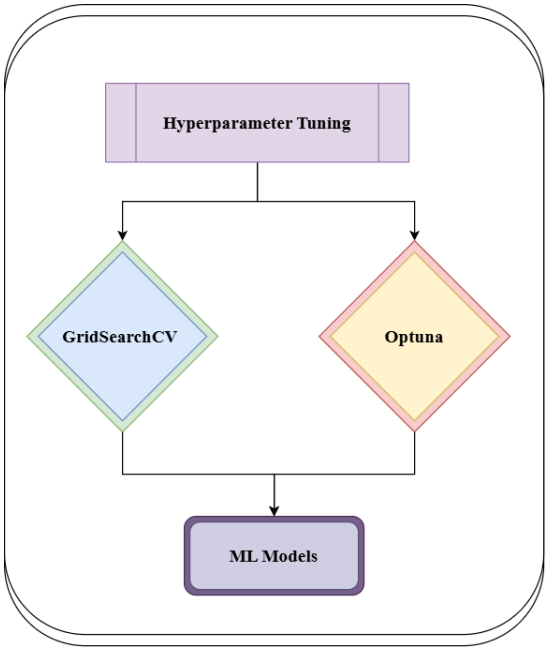

Figure 11: Hyperparameter tuners for the ML models

0.1.6 Applied Transformer Models(LLMs)

To enhance ADHD prediction through natural language understanding, several BERT-based transformer models were employed. These models process medical and survey text data with high accuracy and contextual awareness.

BERT and Variants

BERT (Bidirectional Encoder Representations from Transformers) [22] serves as the foundation for language understanding, pre-trained on 3.3B words from Wikipedia and BooksCorpus. Its lighter variant, DistilBERT [23], retains 97

ClinicalBERT [24] and BioBERT [25] are domain-specific adaptations fine-tuned for clinical and biomedical texts, improving tasks like NER, relation extraction, and medical classification. ALBERT [26] further reduces parameter size via matrix factorization and cross-layer sharing, improving efficiency for large-scale NLP tasks.

LLMs Workflow

As illustrated in Figure 12, the processed dataset was further refined for LLMs by performing tokenization, text encoding, and data splitting. This setup enables BERT-based models to extract meaningful patterns for ADHD-related predictions.

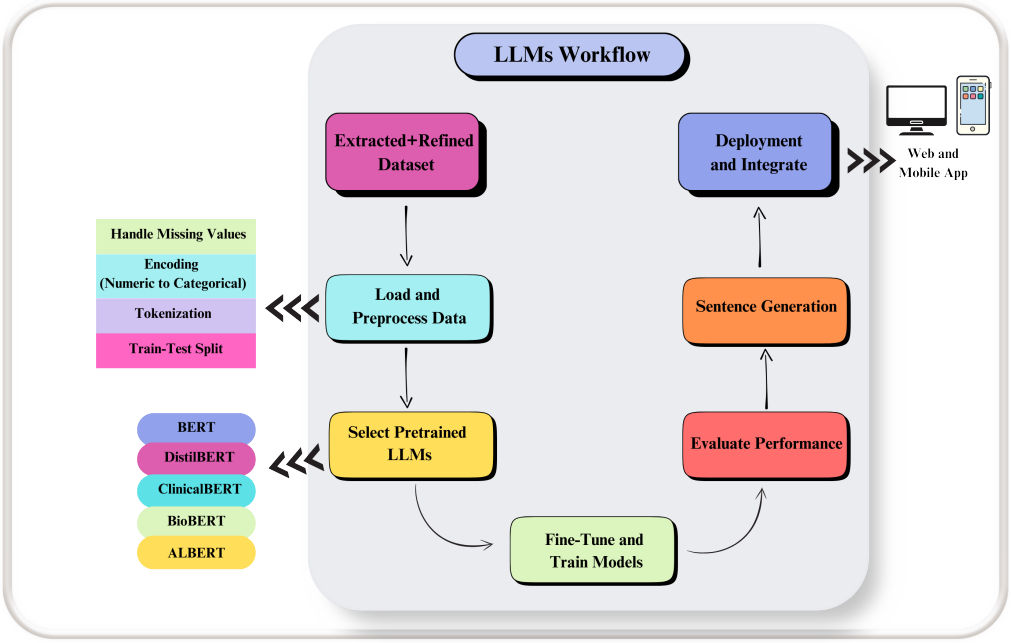

Figure 12: LLMs workflow

The workflow follows a structured pipeline designed to ensure both performance and practical deployability. The process begins with thorough data preparation, including handling missing values, encoding categorical features, performing text tokenization, and splitting the data into training and testing sets. This cleaned and processed dataset is then used to load into selected pre-trained LLMs such as BERT, DistilBERT, ClinicalBERT, RoBERTa, and ALBERT. These models are chosen based on their effectiveness in understanding contextualized health-related text. Once loaded, the models are fine-tuned and trained using the refined dataset. After training, their performance is evaluated using metrics such as accuracy, inference time, and generalization ability. Finally, the best-performing models are integrated into the system for deployment, enabling sentence generation and real-time ADHD-related assistance through both web and mobile applications. All models were evaluated for accuracy, inference time, and generalization. DistilBERT outperformed other variants with the highest accuracy and fastest runtime demonstrating its practical advantage for deployment.

0.1.7 Dialogflow-based therapist Chatbot "Comfort Chat"

The proposed system includes a lightweight and user-friendly chatbot named ComfortChat, developed using Google's Dialogflow platform [27]. Designed to offer primary therapeutic support, ComfortChat enhances user engagement by delivering quick, relevant responses.

Dialogflow is widely adopted in healthcare for building virtual assistants, symptom check-

ers, and patient support tools. It leverages natural language understanding (NLU) to identify medical intents and extract key entities such as symptoms and conditions. Its integration capabilities, multilingual support, and compliance with healthcare standards make it ideal for real-time medical interaction.

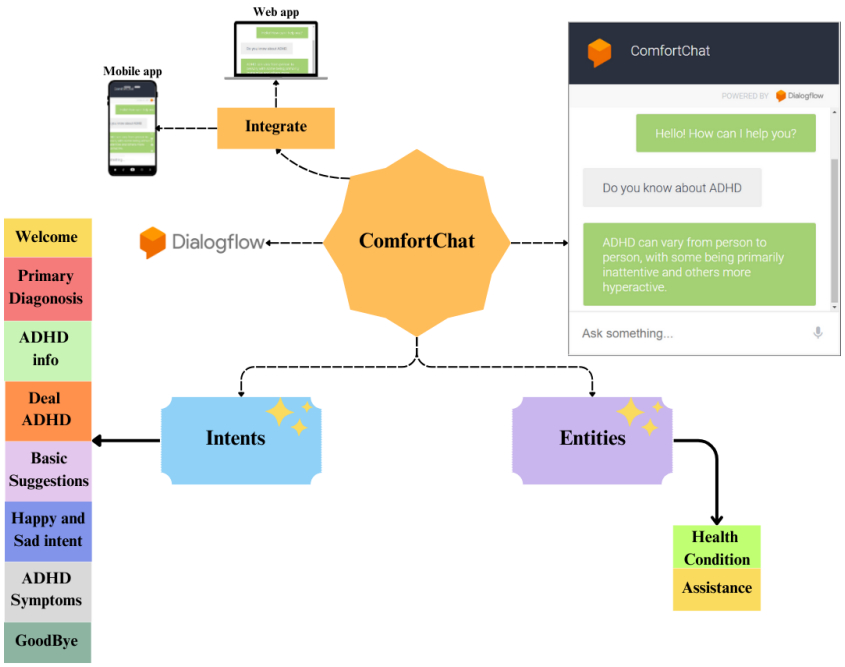

Figure 13: Structure of Chatbot (ComfortChat)

Structure of ComfortChat

ComfortChat is built using Dialogflow and integrated with both web and mobile apps for user accessibility. The system is driven by two core components: Intents and Entities.

- **Intents:** Guide responses based on user queries. Examples include: Welcome, Diagnosis, ADHD Info, Suggestions, Emotions, and Goodbye.
- **Entities:** Extract key data like *Health Condition* and *Assistance* to make replies more relevant.

This structure enables ComfortChat to deliver helpful, context-aware support for ADHD-related concerns across platforms.

The overall pipeline spans from data collection to deployment, incorporating rigorous processing at every stage.

Bibliography

[1] Child and Adolescent Health Measurement Initiative (CAHMI). 2021-2022 national survey of children's health (nsch) indicator dataset. Data Resource Center for Child and Adolescent Health, supported by HRSA, MCHB, 2025. Retrieved February 24, 2025, from `www.childhealthdata.org`.

[2] Johns Hopkins Bloomberg School of Public Health. Child and adolescent health measurement initiative (cahmi). `https://www.jhsph.edu/cahmi`, 2023.

[3] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2201–2206. ACM, June 2016.

[4] W. McKinney et al. pandas: Powerful data analysis toolkit. `https://pandas.pydata.org/`, 2022. Accessed: August 6, 2022.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3):52, 2021.

[7] S. Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, 24(1):12–18, 2014.

[8] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc., Sebastopol, CA, 2017.

[9] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[10] S. Chen, G. I. Webb, L. Liu, and X. Ma. A novel selective naïve bayes algorithm. *Knowledge-Based Systems*, 192:105361, 2020.

[11] M. Awad and R. Khanna. *Support Vector Machines for Classification*. A Press, Berkeley, CA, 2015.

[12] M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[13] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[14] B. Charbuty and A. Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1):20–28, March 2021.

[15] A. K. Singh. Breast cancer classification using ml on wdbc. In *Lecture Notes in Electrical Engineering*, pages 609–619. January 2023.

[16] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[17] J. Zhang, D. Mucs, U. Norinder, and F. Svensson. Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets. *Journal of Chemical Information and Modeling*, 59(10):4150–4158, 2019.

[18] J. T. Hancock and T. M. Khoshgoftaar. Catboost for big data: An interdisciplinary review. *Journal of Big Data*, 7:94, 2020.

[19] B. Pavlyshenko. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 255–258, 2018.

[20] Nicos Isaak and Loizos Michael. Blending nlp and machine learning for the development of winograd schemas. 03 2021.

[21] R. U. Mhapsekar, N. O'Shea, S. Davy, and L. Abraham. Hybrid blended deep learning approach for milk quality analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2253–2268, June 2024.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805, 2018.

[23] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint*, arXiv:1909.10351, 2019.

[24] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint*, arXiv:1904.05342, 2019.

[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint*, arXiv:1909.11942, 2020.

[27] A. F. Muhammad, D. Susanto, A. Alimudin, F. Adila, M. H. Assidiqi, and S. Nabhan. Developing english conversation chatbot using dialogflow. In *2020 International Electronics Symposium (IES)*, pages 468–475. IEEE, September 2020.