The 3rd International Conference on Intelligence of Things 2024
(ICITconf'2024)

ADDRESSING DATA IMBALANCE IN INSURANCE FRAUD PREDICTION USING SAMPLING TECHNIQUES AND ROBUST LOSSES

Tai Do Nhu, Loc Dinh Tan, Di Khanh Le, Huy Nguyen Quoc

donhutai@gmail.com, locdinh.31221020226@st.ueh.edu.vn,
khanhle.31211022918@st.ueh.edu.vn, nqhuy@sgu.edu.vn

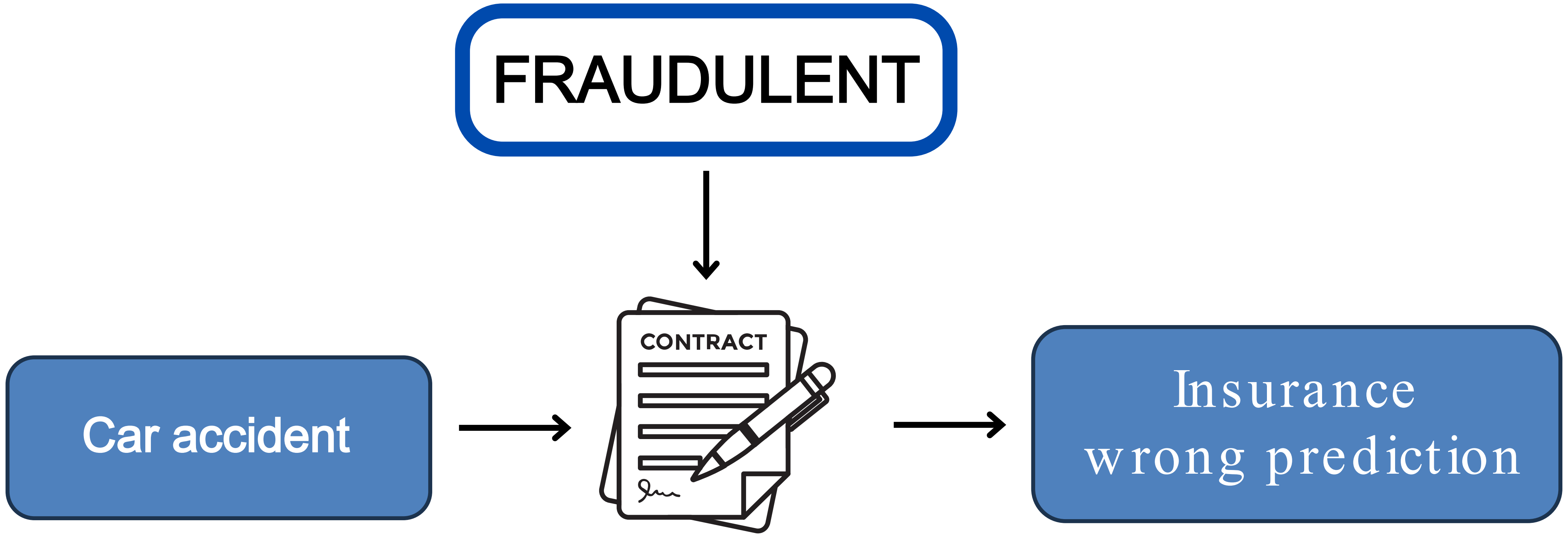
September 2024

AGENDA

1. Introduction
2. Motivation
3. Related works
4. Proposed method
5. Experiment and Results
6. Conclusion

1. INTRODUCTION

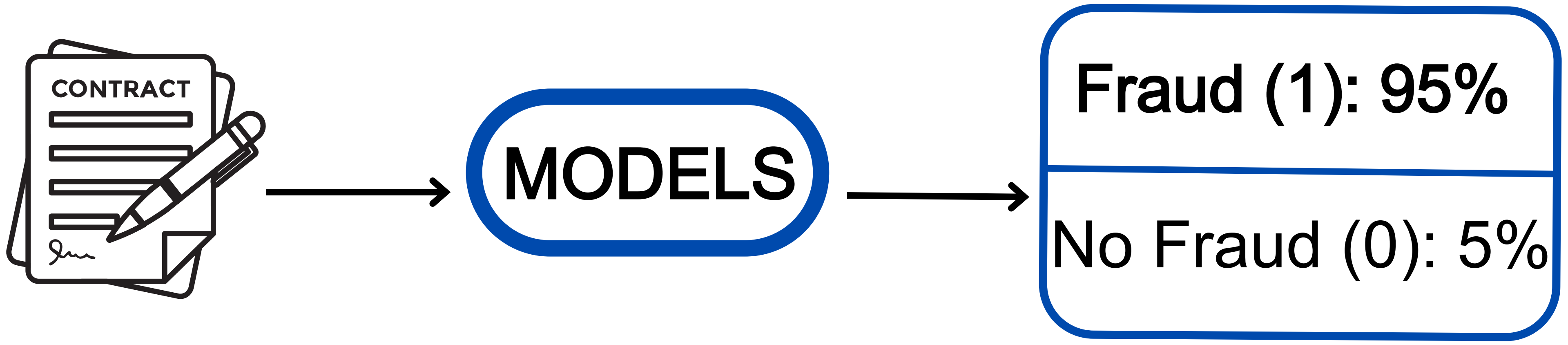
The situation:



The problem:

Input:
Features in contracts

Output:
Probabilities off binary classes



1. INTRODUCTION

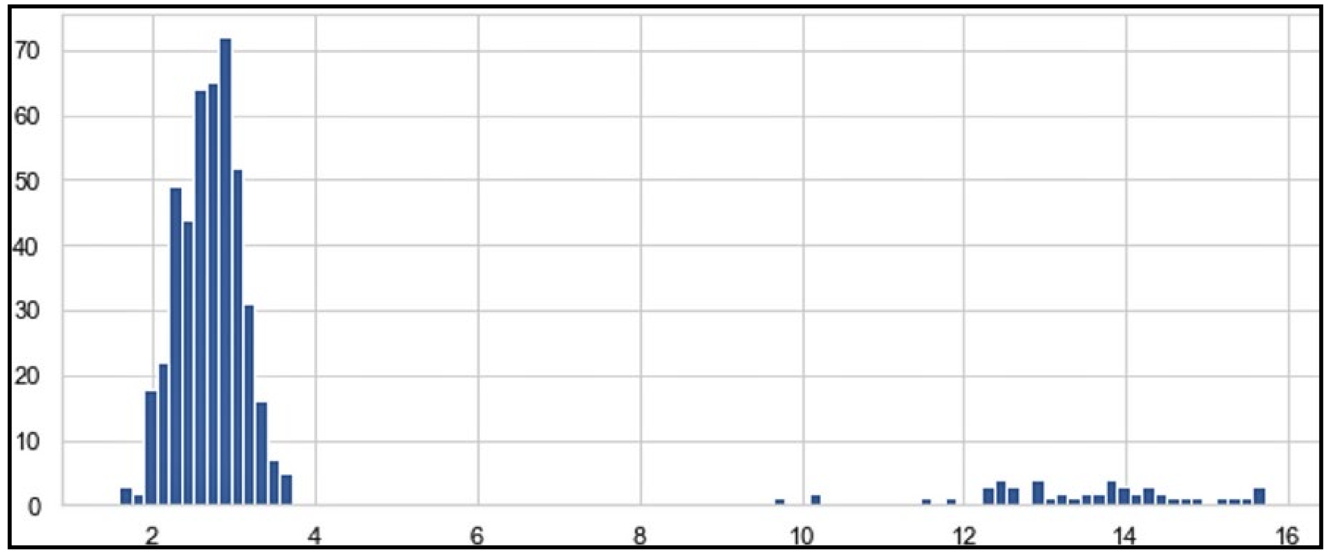
- **Fraudsters often provide false information to claim insurance money.**
- The provided data often has a range of hidden issues.

FRAUDULENT

Fake Figures

Missing data

Fake validation

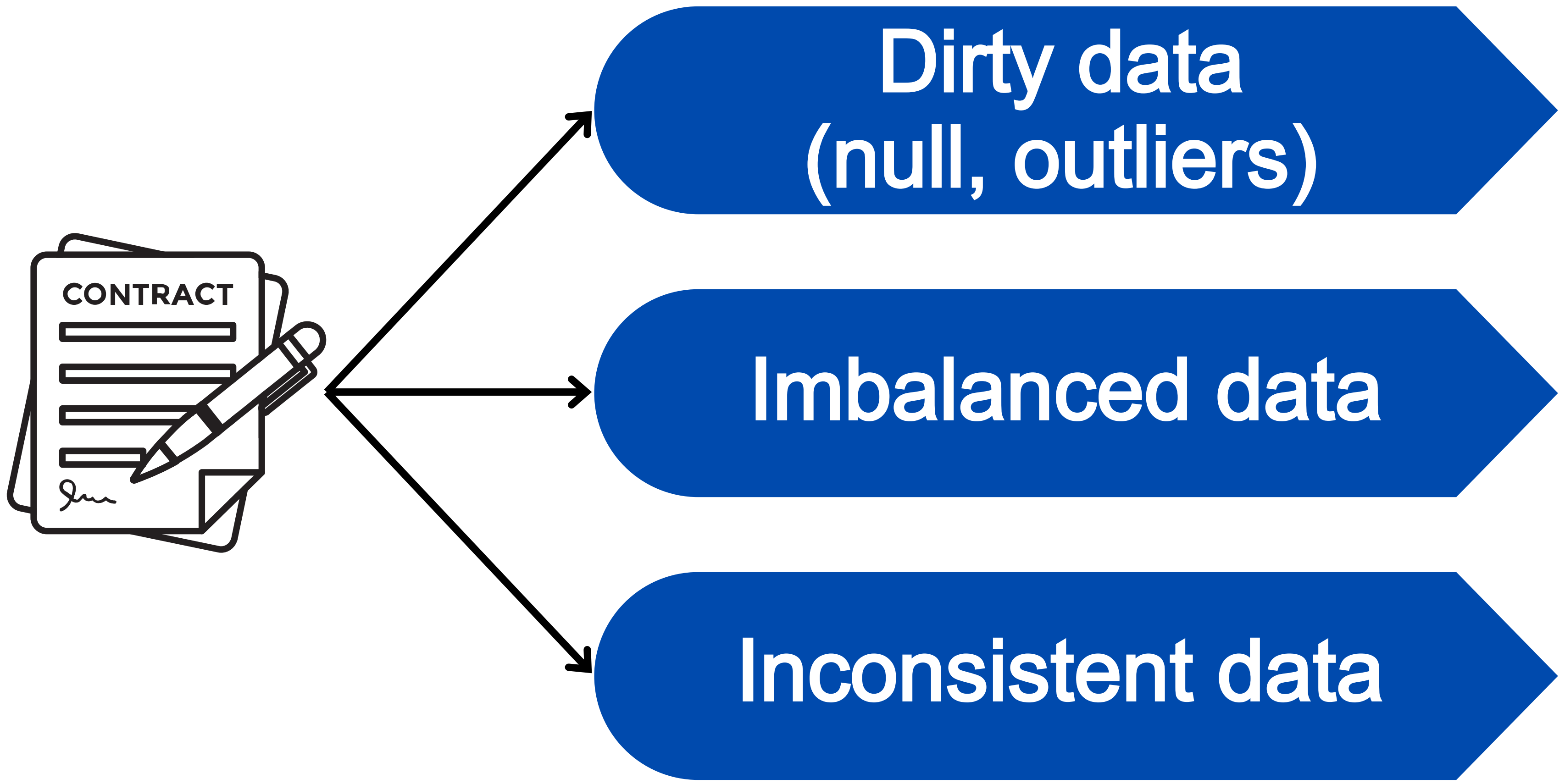


Claimants often fail to provide or are slow to supply sufficient information.

Fraudsters inflate claims using fake documents or insurance employee connections.

1. INTRODUCTION

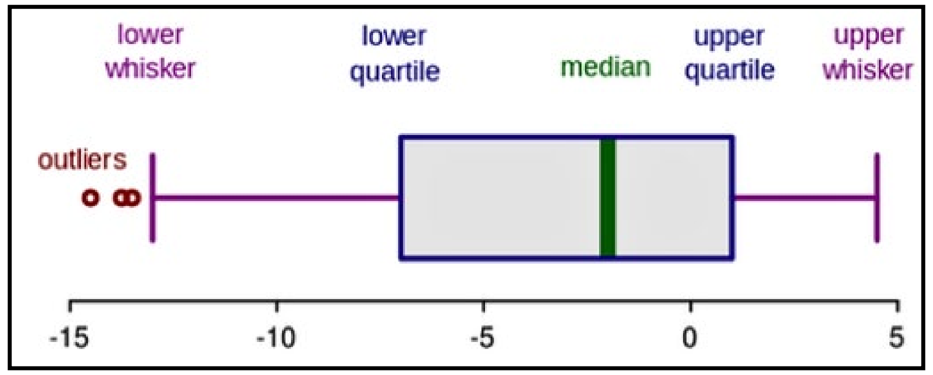
- Fraudsters often provide false information to claim insurance money.
- **The provided data often has a range of hidden issues.**

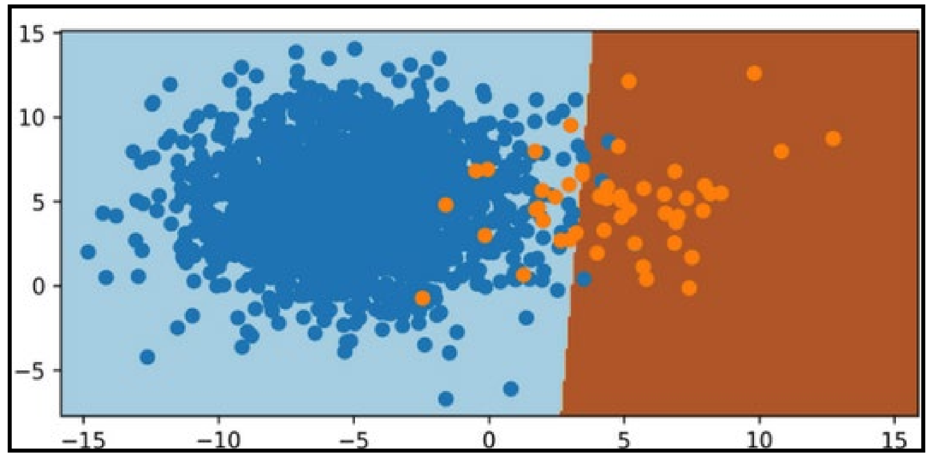


Dirty data (null, outliers)

Imbalanced data

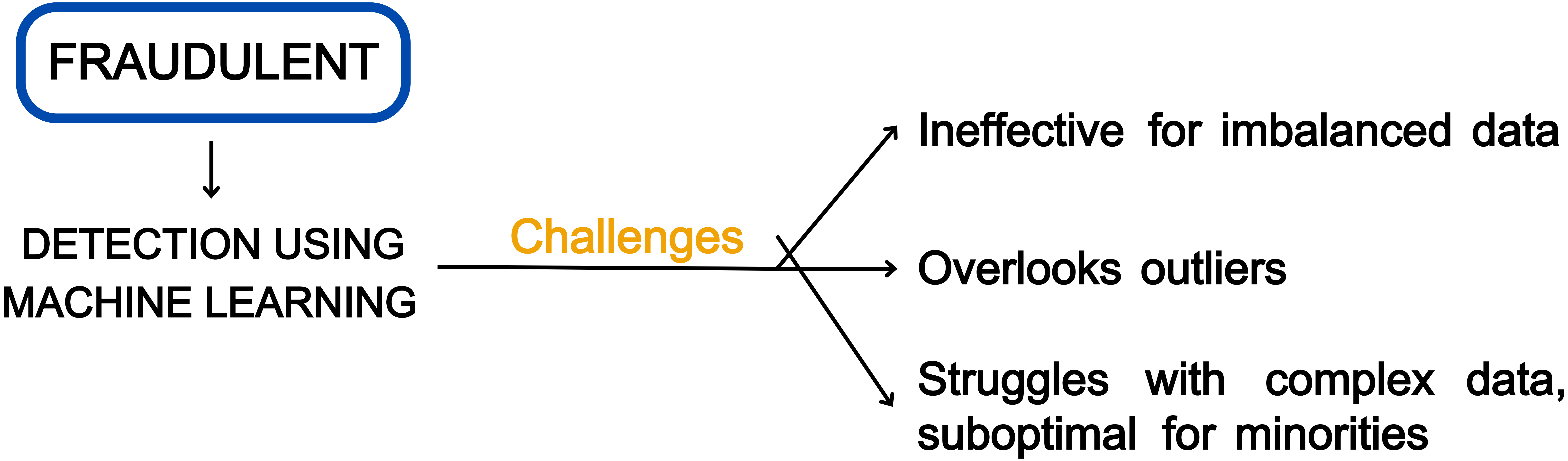
Inconsistent data





Car insurance claims fluctuate due to delays in information and procedures.

1. INTRODUCTION



FRAUDULENT

DETECTION USING MACHINE LEARNING

Challenges

Ineffective for imbalanced data

Overlooks outliers

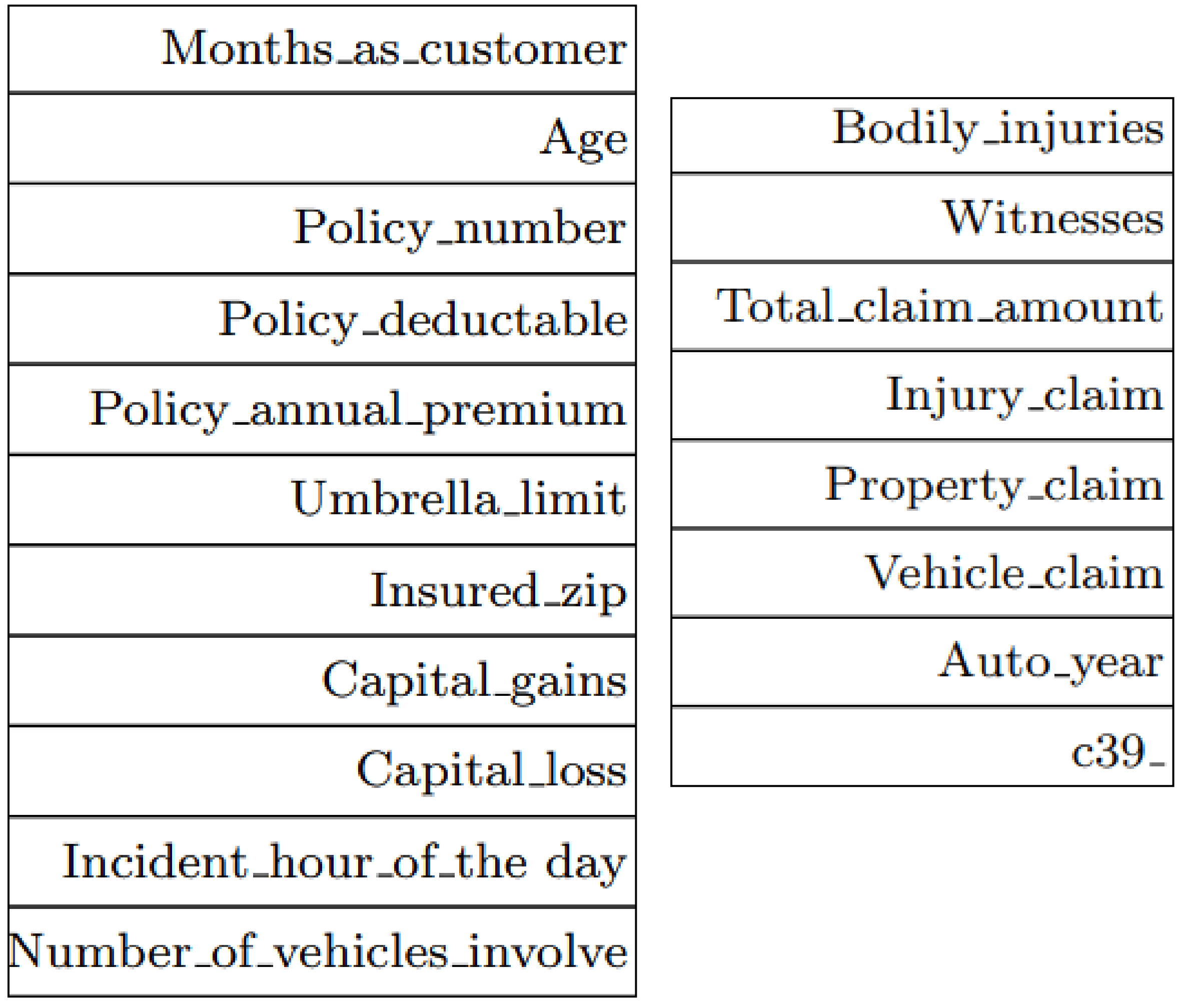
Struggles with complex data, suboptimal for minorities

How can deep learning models be improved to effectively handle imbalanced data?

1. INTRODUCTION

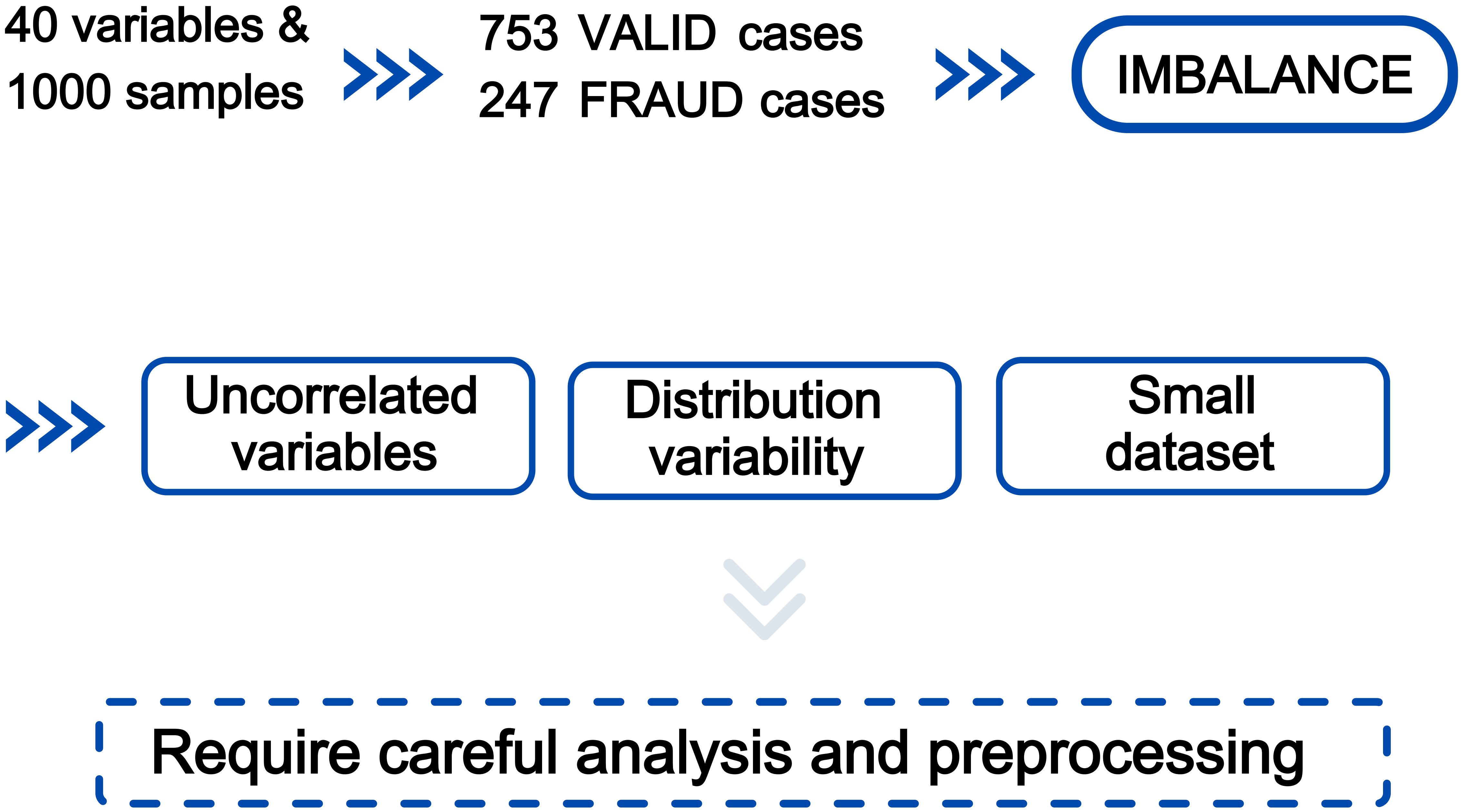
Dataset

Source: Kaggle [1]



Months_as_customer
Age
Policy_number
Policy_deductable
Policy_annual_premium
Umbrella_limit
Insured_zip
Capital_gains
Capital_loss
Incident_hour_of_the day
Number_of_vehicles_involve

Bodily_injuries
Witnesses
Total_claim_amount
Injury_claim
Property_claim
Vehicle_claim
Auto_year
c39_



40 variables &
1000 samples

753 VALID cases
247 FRAUD cases

IMBALANCE

Uncorrelated variables

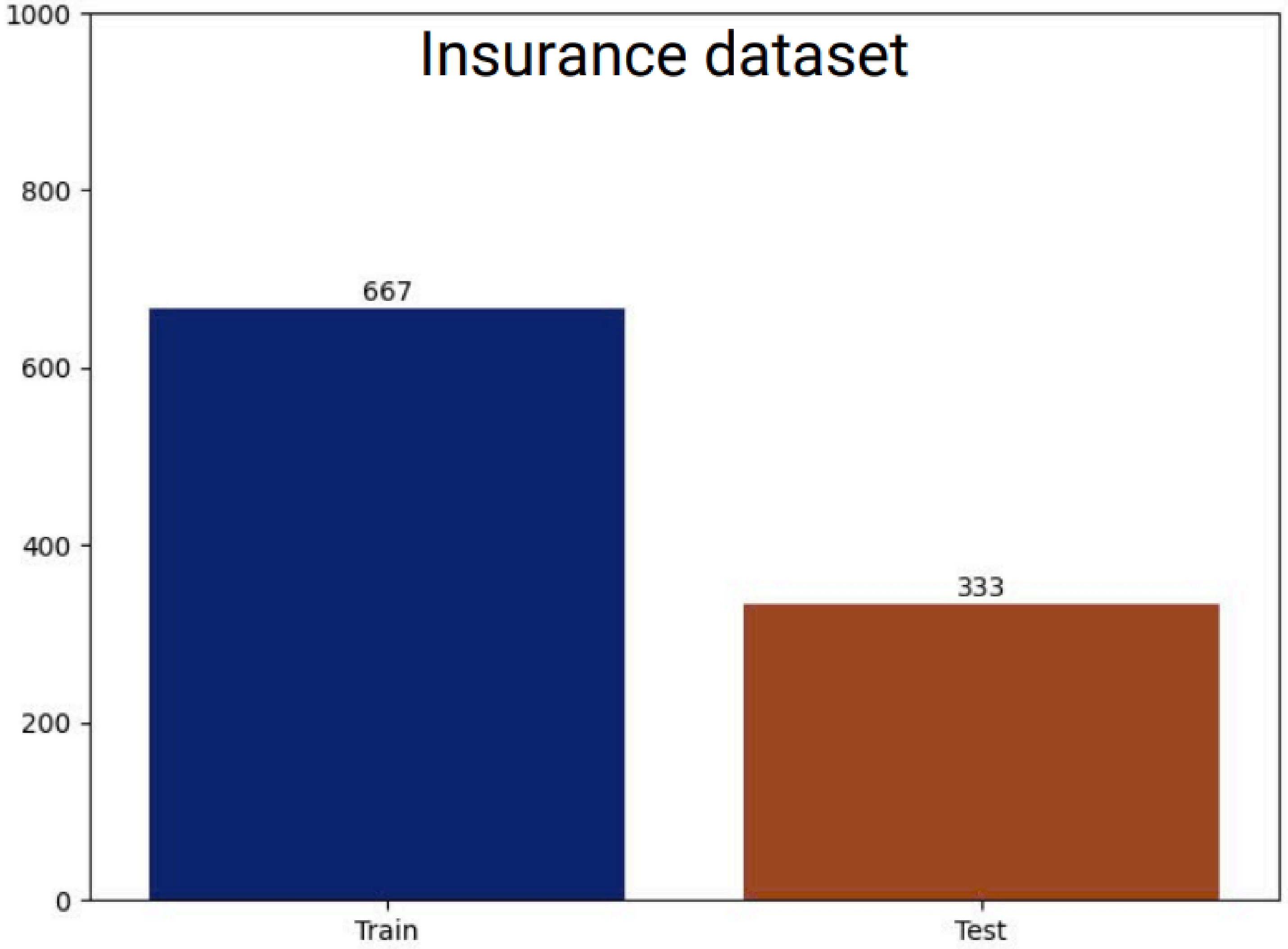
Distribution variability

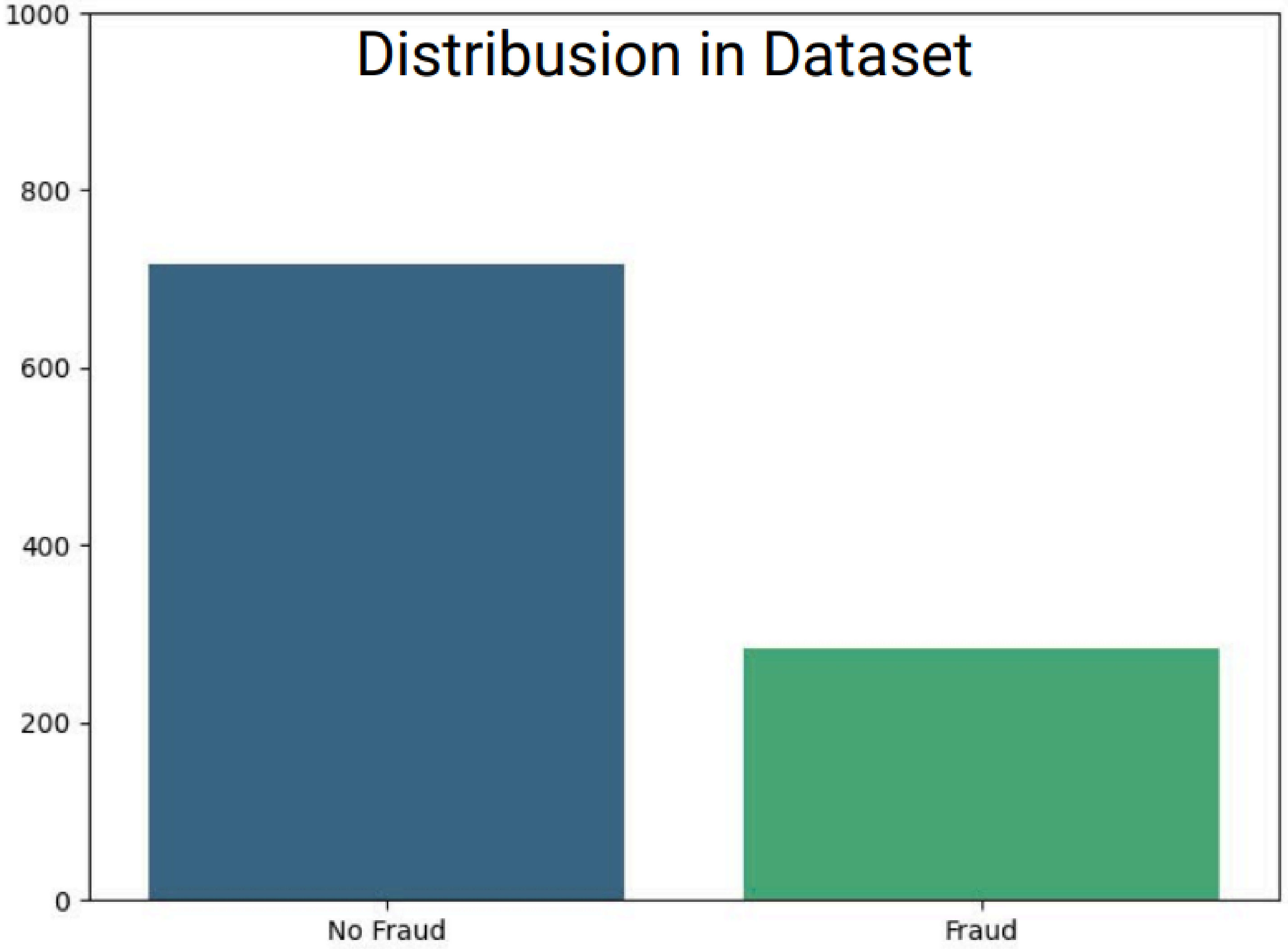
Small dataset

Require careful analysis and preprocessing

[1] Jhamtani,A.(n.d.).Automobile insurance.(2018,December 27).Kaggle.

2. MOTIVATION





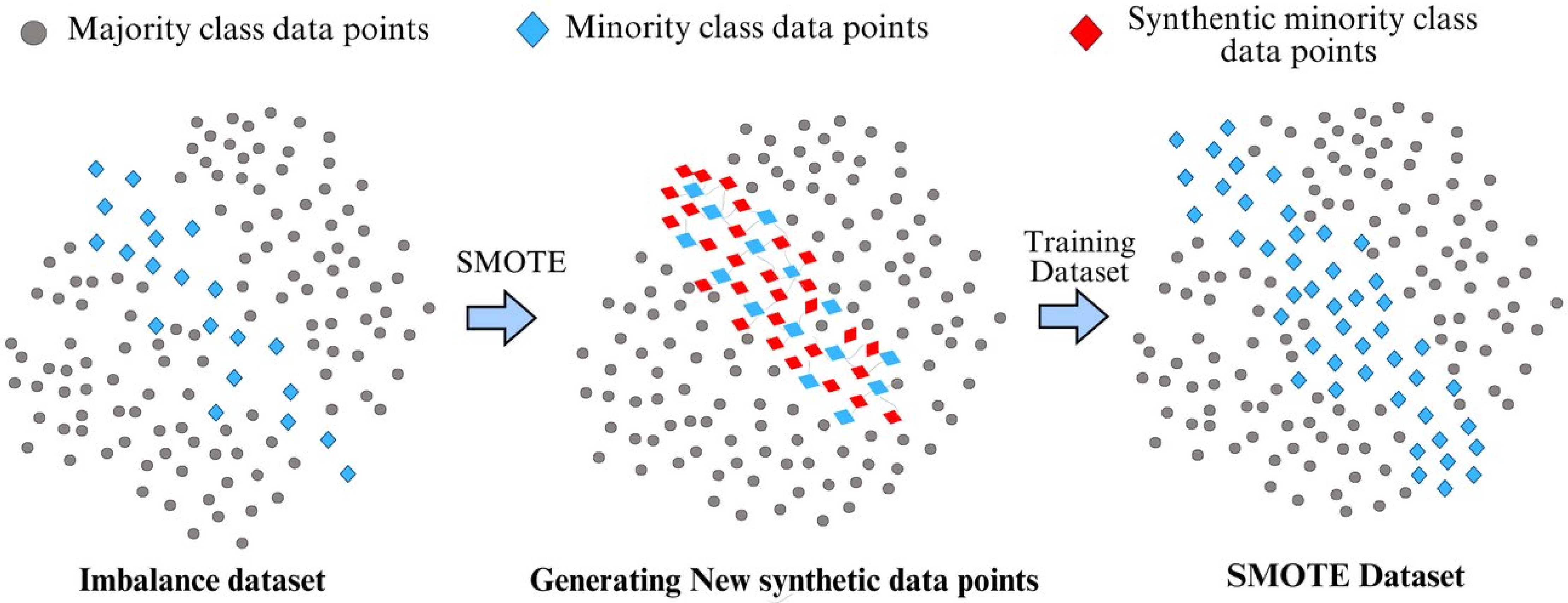
Propose deep learning models:

- Address data imbalance.
- Enhance performance with robust sampling and loss functions.

3. RELATED WORKS

- **Dablain et al[1]**

SMOTE can be used to address issues caused by **imbalanced data** by **generating new observation points from the original data**, which helps **the model learn and classify more effectively**.

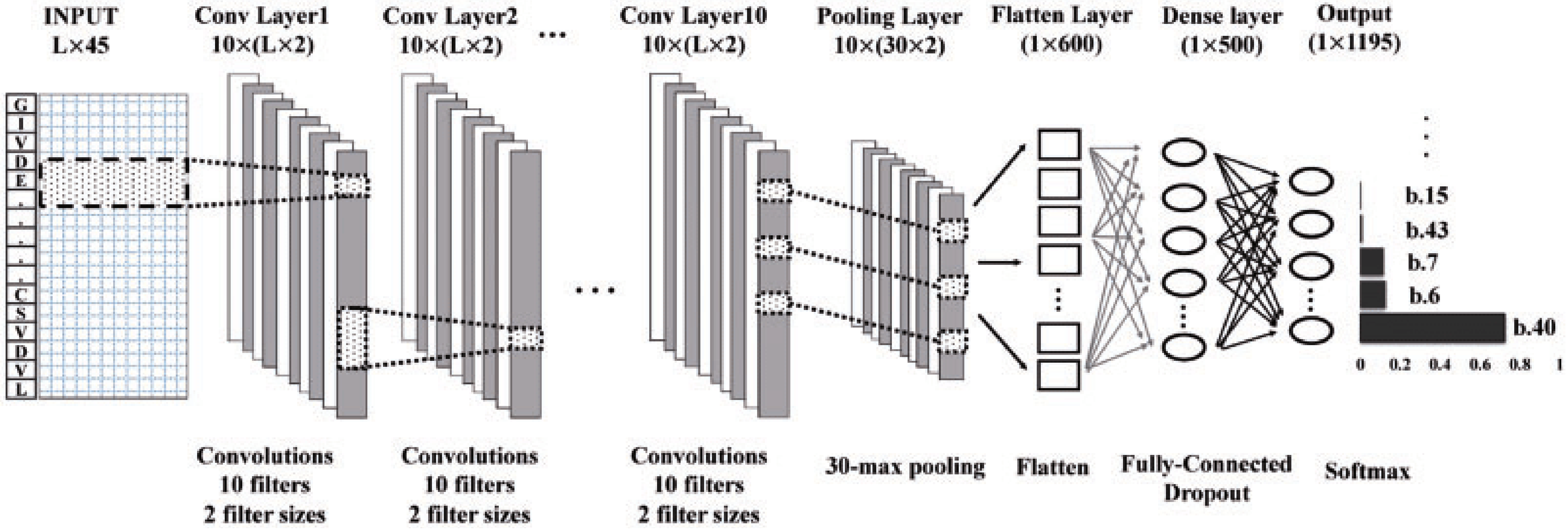

● Majority class data points
Minority class data points
Synthetic minority class data points
SMOTE
Training Dataset
Imbalance dataset
Generating New synthetic data points
SMOTE Dataset

[1] Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). **DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data.** IEEE Transactions on Neural Networks and Learning Systems, 34(9), 6390-6404.

3. RELATED WORKS

- **Azizjon et al1**

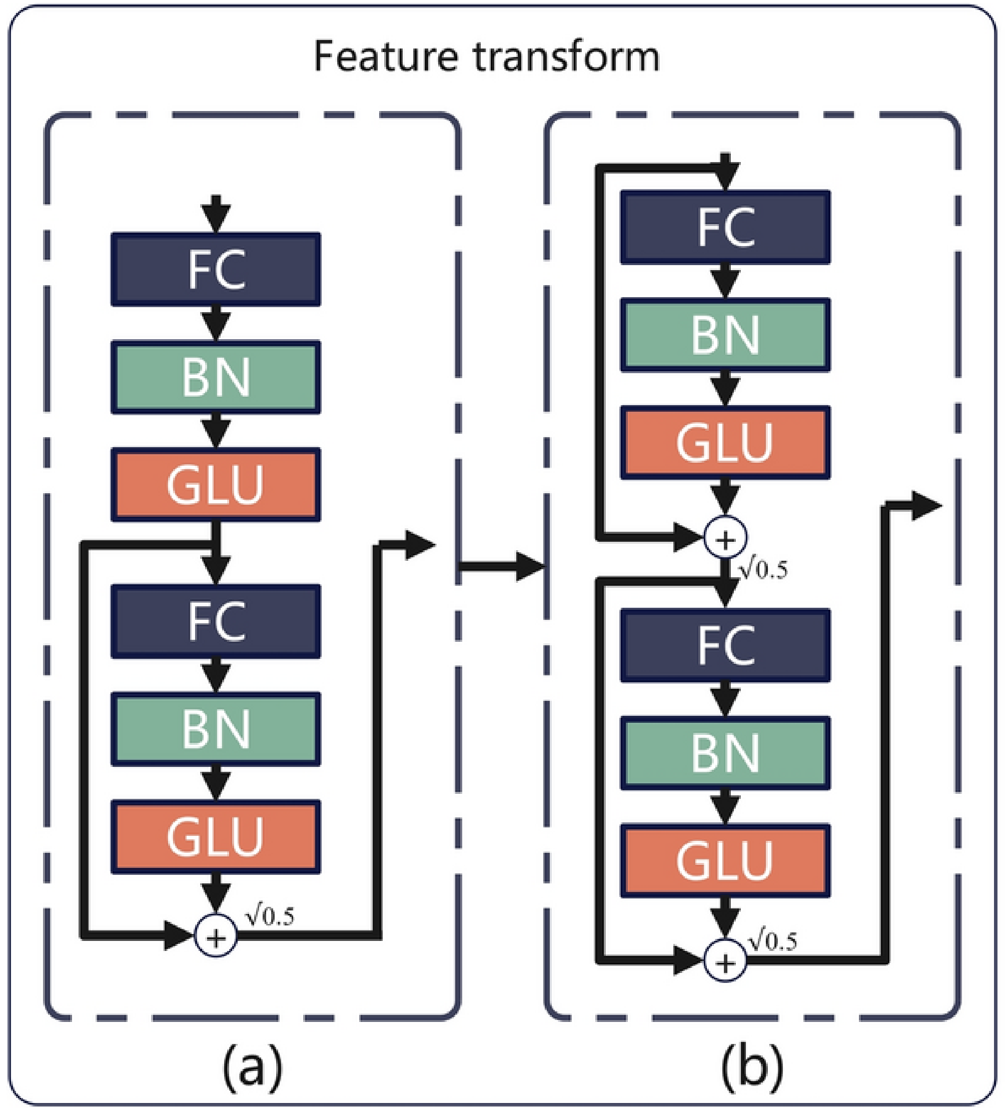
The **Convolutional Neural Network (CNN) model,** renowned for **image processing tasks,** leverages its convolutional architecture, allowing it to effectively extract information from input data. The application of CNN architectures to t**abular data problems** has demonstrated **significant efficiency, extending its effectiveness beyond just image-related fields.**

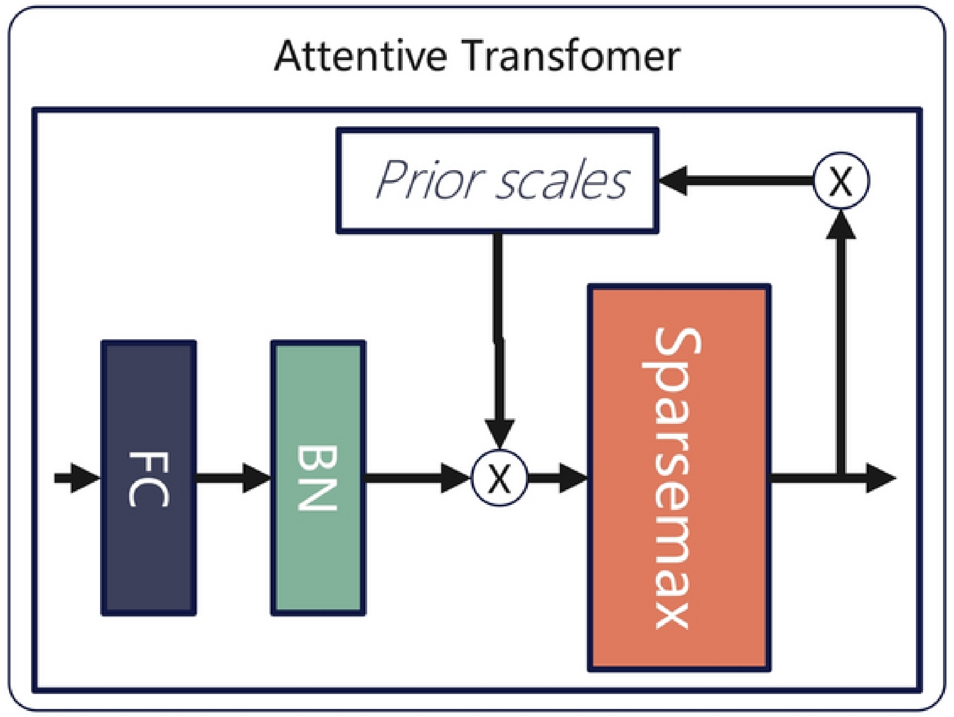


[1] Azizjon, M., Jumabek, A., & Kim, W. (2020, February). **1D CNN based network intrusion detection with normalization on imbalanced data**. In 2020 international conference on artificial intelligence in information and communication (ICAIIC) (pp. 218-224). IEEE.

3. RELATED WORKS

- **Arik et al[1]**





TabNet is **a deep learning architecture** specifically designed for **tabular data**. It utilizes a **sequential attention mechanism**, enabling the model to select relevant features dynamically at each decision step, leading to **high accuracy and robust classification performance** in various **tabular data tasks**.

[1] Arik, S. Ö., & Pfister, T. (2021, May). **Tabnet: Attentive interpretable tabular learning**. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 8, pp. 6679-6687).

3. RELATED WORKS

- **Losses**

F1 loss= $1 - \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$

where h = (h1,..,hm) ∈ {0,1} is a prediction of an m-dimensional binary label vector y=(y1,..,ym) (e.g., the class labels of a test set of size m in binary)

Focal loss= $-(1 - p_t)^{\gamma} \log(p_t)$

where pt is the predicted probability for the target class and γ is a focusing parameter.

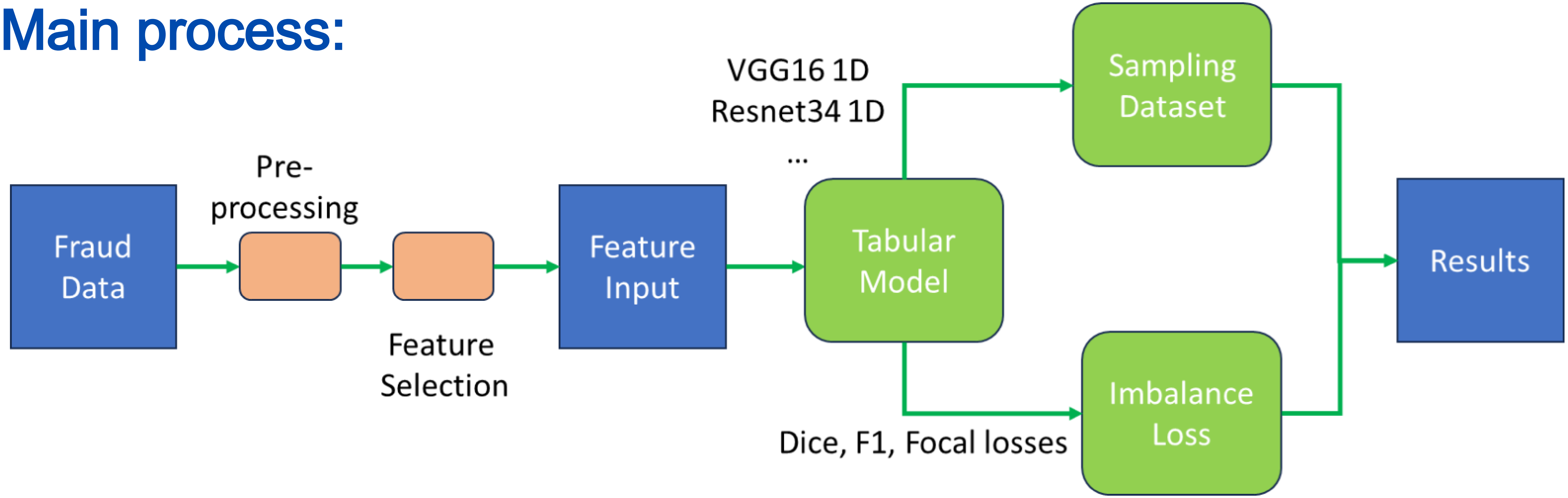
Dice loss= $1 - \frac{2 \times |X \cap Y|}{|X| + |Y|}$

where X and Y are the predicted and ground truth sets.

4. PROPOSED METHOD

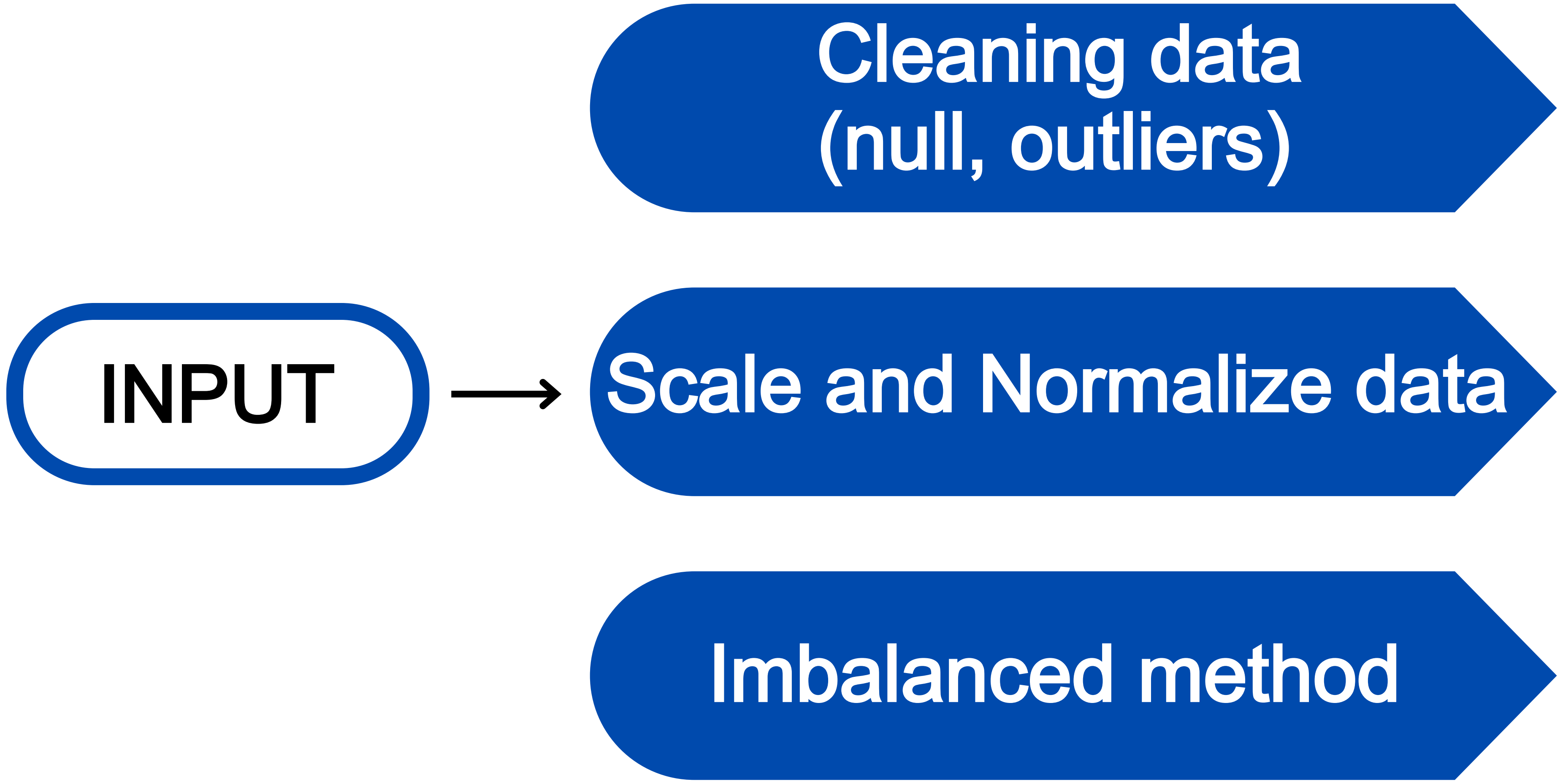
- Build a model using tabular models for this problem (Assess model stability on small datasets).
- Use Sampling Dataset and Imbalance Loss approach on our proposed model.

Main process:



4. PROPOSED METHOD

Explanation Detailed: Dataset preprocesing



INPUT

Cleaning data (null, outliers)

Scale and Normalize data

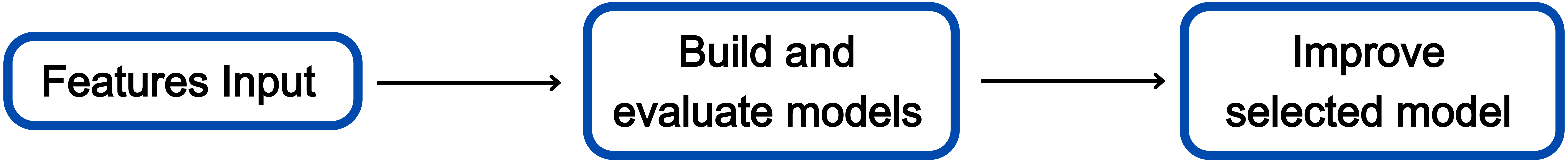
Imbalanced method

- Using Standard scaler for numeric feature
- Using Label encoder for categorical feature
- White balncing
- SMOTE

4. PROPOSED METHOD

Explanation Detailed: Build and evaluate base model

- Use feature selection techniques corresponding to each model to optimize performance.
- Choose best tabular models and improve models performance.



Features Input

Build and evaluate models

Improve selected model

- Feature selection algorithm: RFE, Lasso, correlation

- Evaluate by Cross-validation
- Choose best models
- Using evaluation metrics

- Hyparameter tuning with Gridsearch CV

4. PROPOSED METHOD

Models:

built an MLP with an input layer, shallow Dense layers using ReLU & SoftMax classification output.

using TabNet to assess how data imbalance impacts its specialized capabilities.

redesign CNN architectures with 1D convolution and pooling to evaluate models like VGG16, ResNet, and Inception.

4. PROPOSED METHOD

Models:

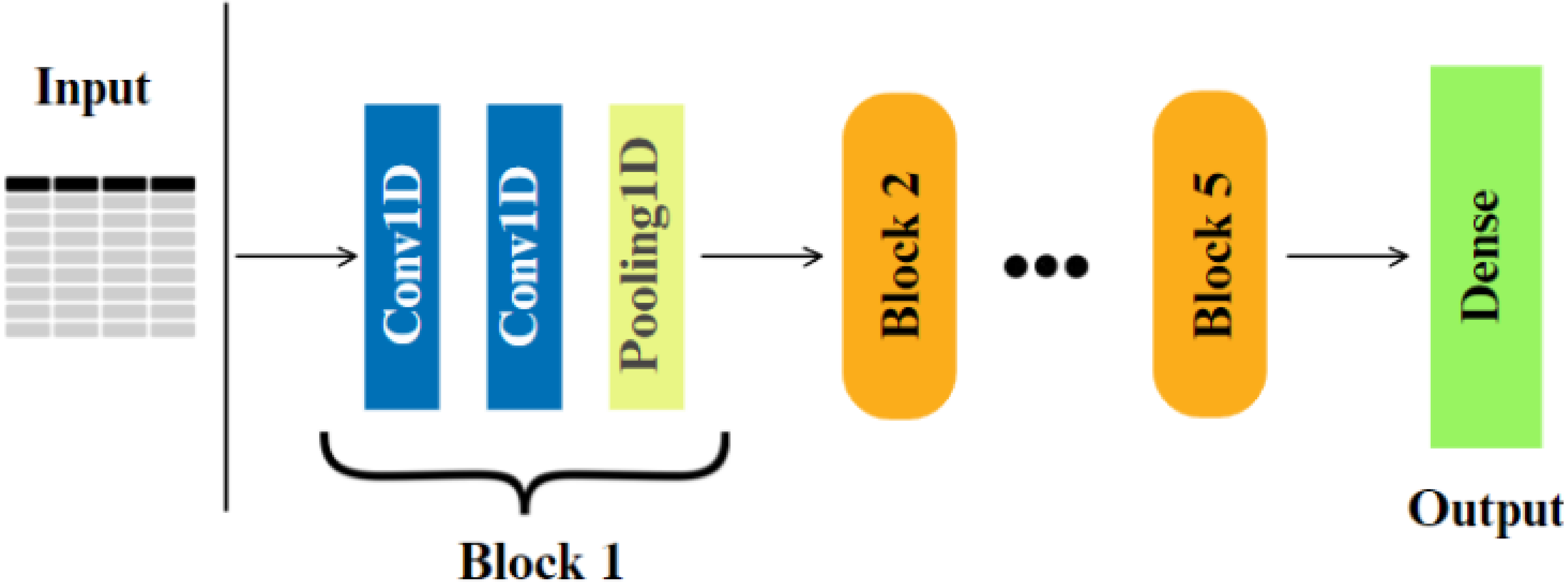


Fig. 2: Model architecture designed based on VGG16 for tabular data.

4. PROPOSED METHOD

Loss functions:

F1 LOSS

$$\text{F1 Loss} = 1 - \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Balances precision and recall for imbalanced data, suitable for detecting positive and negative instances.

FOCAL LOSS

$$\text{Focal Loss} = -(1 - p_t)^{\gamma} \log(p_t)$$

Addresses data imbalance by focusing on harder samples, mitigating bias towards dominant classes.

DICE LOSS

$$\text{Dice Loss} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

Originally used in medical imaging, applied to imbalanced tabular data to optimize the Dice coefficient.

MULTI LOSS

$$\text{Multi Loss} = \alpha \times \text{Loss}_1 + \beta \times \text{Loss}_2 + \gamma \times \text{Loss}_3$$

Combines multiple loss functions to leverage their unique advantages and minimize individual weaknesses.

4. PROPOSED METHOD

Loss functions:

F1 LOSS

$$\text{F1 Loss} = 1 - \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Balances precision and recall for imbalanced data, suitable for detecting positive and negative instances.

FOCAL LOSS

$$\text{Focal Loss} = -(1 - p_t)^\gamma \log(p_t)$$

Addresses data imbalance by focusing on harder samples, mitigating bias towards dominant classes.

DICE LOSS

$$\text{Dice Loss} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

Originally used in medical imaging, applied to imbalanced tabular data to optimize the Dice coefficient.

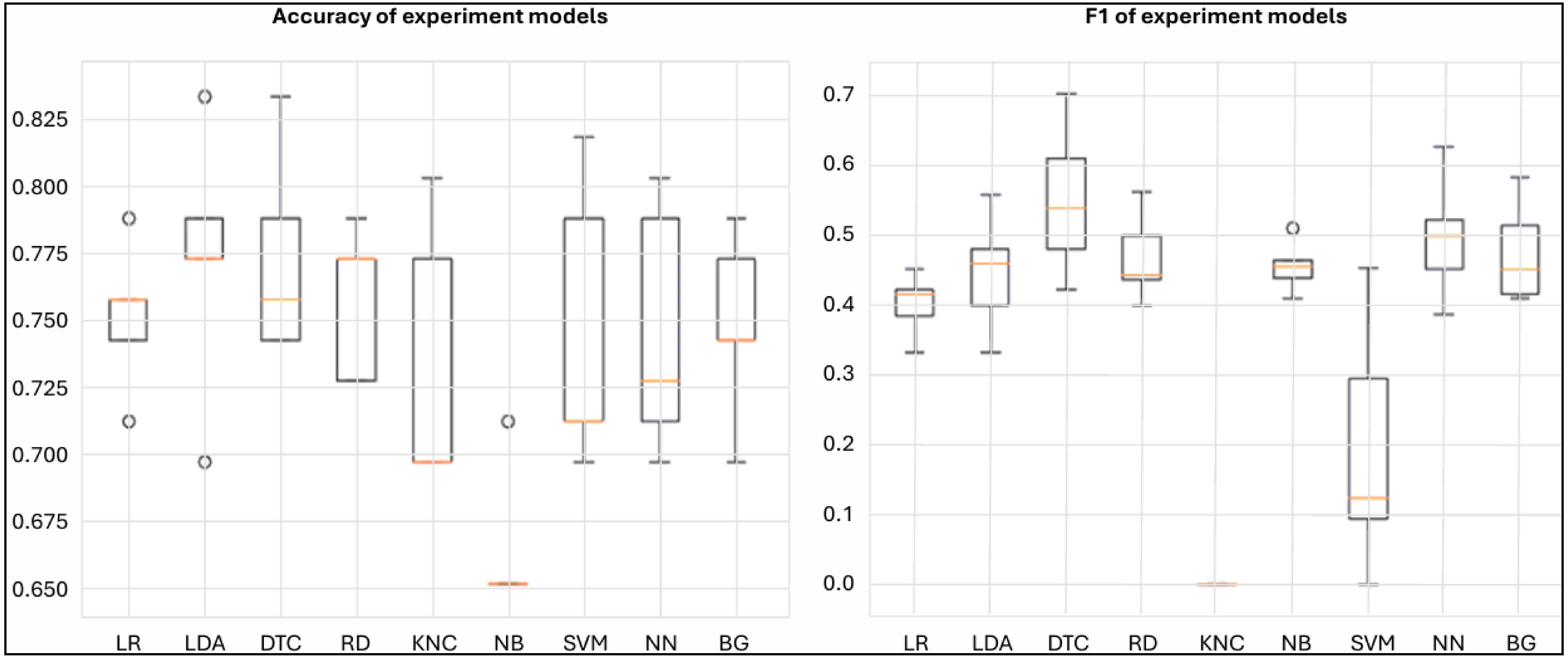
MULTI LOSS

$$\text{Multi Loss} = \alpha \times \text{Loss}_1 + \beta \times \text{Loss}_2 + \gamma \times \text{Loss}_3$$

Combines multiple loss functions to leverage their unique advantages and minimize individual weaknesses.

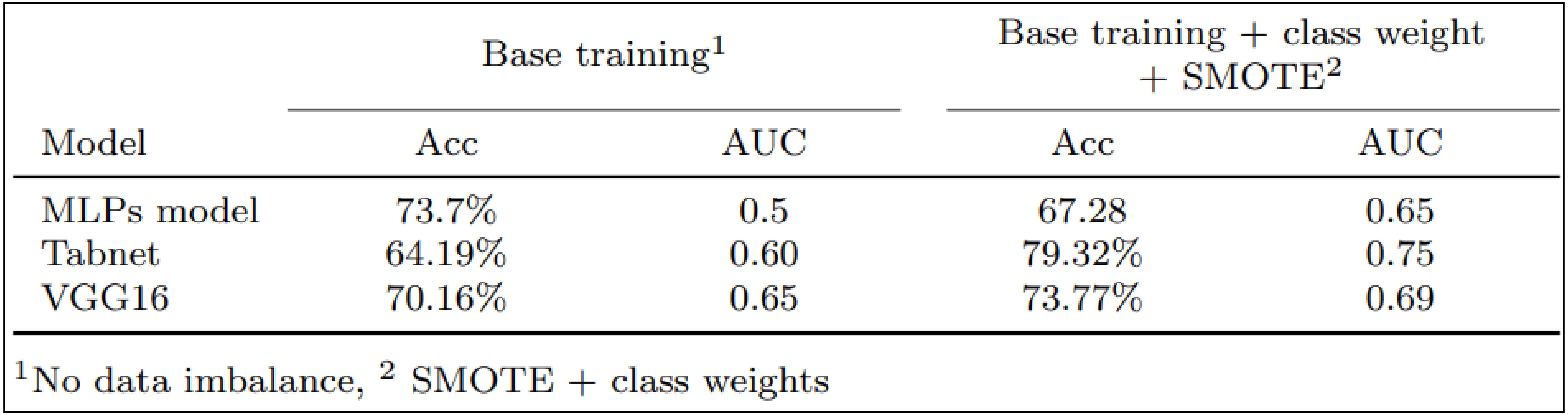
5. EXPERIMENTS AND RESULTS

5.1. Traditional models



5. EXPERIMENTS AND RESULTS

5.2. DEEP LEARNING models

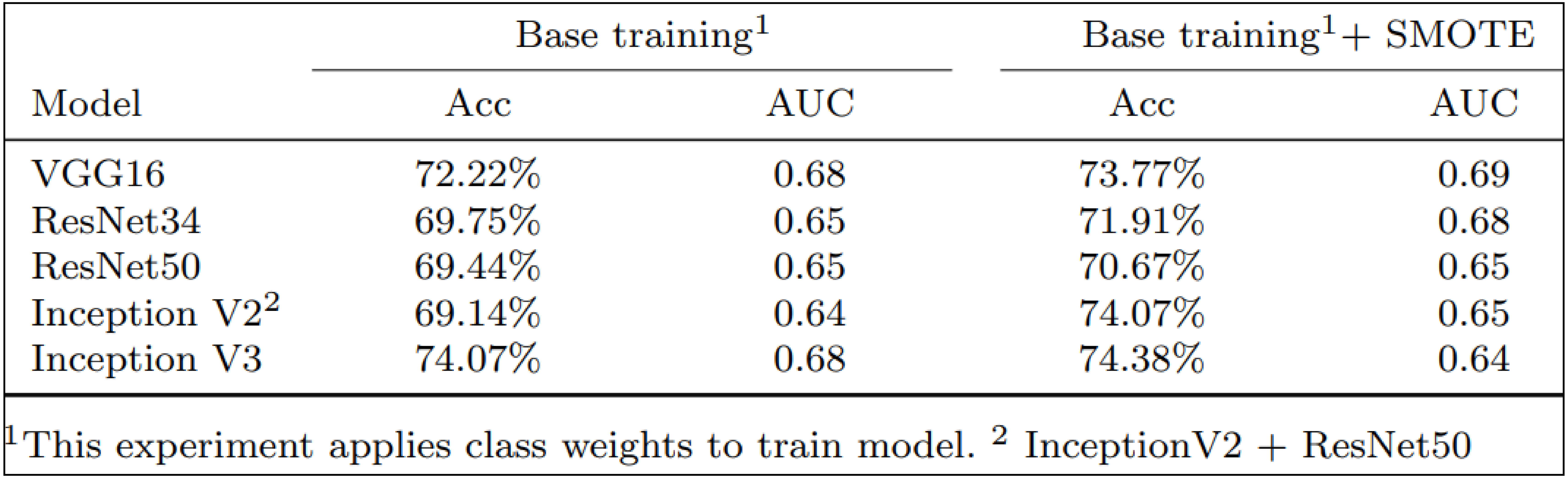


	Base training[1]		Base training + class weight + SMOTE[2]	
Model	Acc	AUC	Acc	AUC
MLPs model	73.7%	0.5	67.28	0.65
Tabnet	64.19%	0.60	79.32%	0.75
VGG16	70.16%	0.65	73.77%	0.69

[1] No data imbalance, [2] SMOTE + class weights

5. EXPERIMENTS AND RESULTS

5.3. CNN models applying imbalance handling strategies

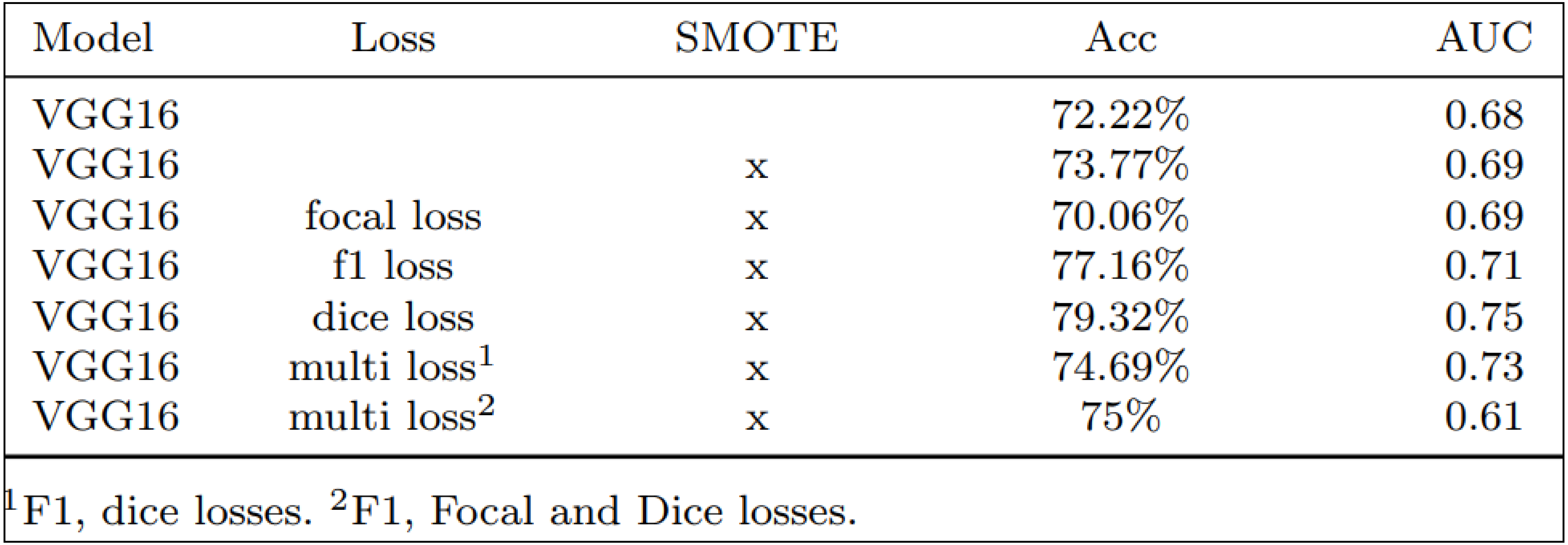


	Base training[1]		Base training[1] + SMOTE	
Model	Acc	AUC	Acc	AUC
VGG16	72.22%	0.68	73.77%	0.69
ResNet34	69.75%	0.65	71.91%	0.68
ResNet50	69.44%	0.65	70.67%	0.65
Inception V2[2]	69.14%	0.64	74.07%	0.65
Inception V3	74.07%	0.68	74.38%	0.64

[1] This experiment applies class weights to train model. [2] InceptionV2 + ResNet50

5. EXPERIMENTS AND RESULTS

5.4. Accuracy and AUC of VGG16 models through situationsies

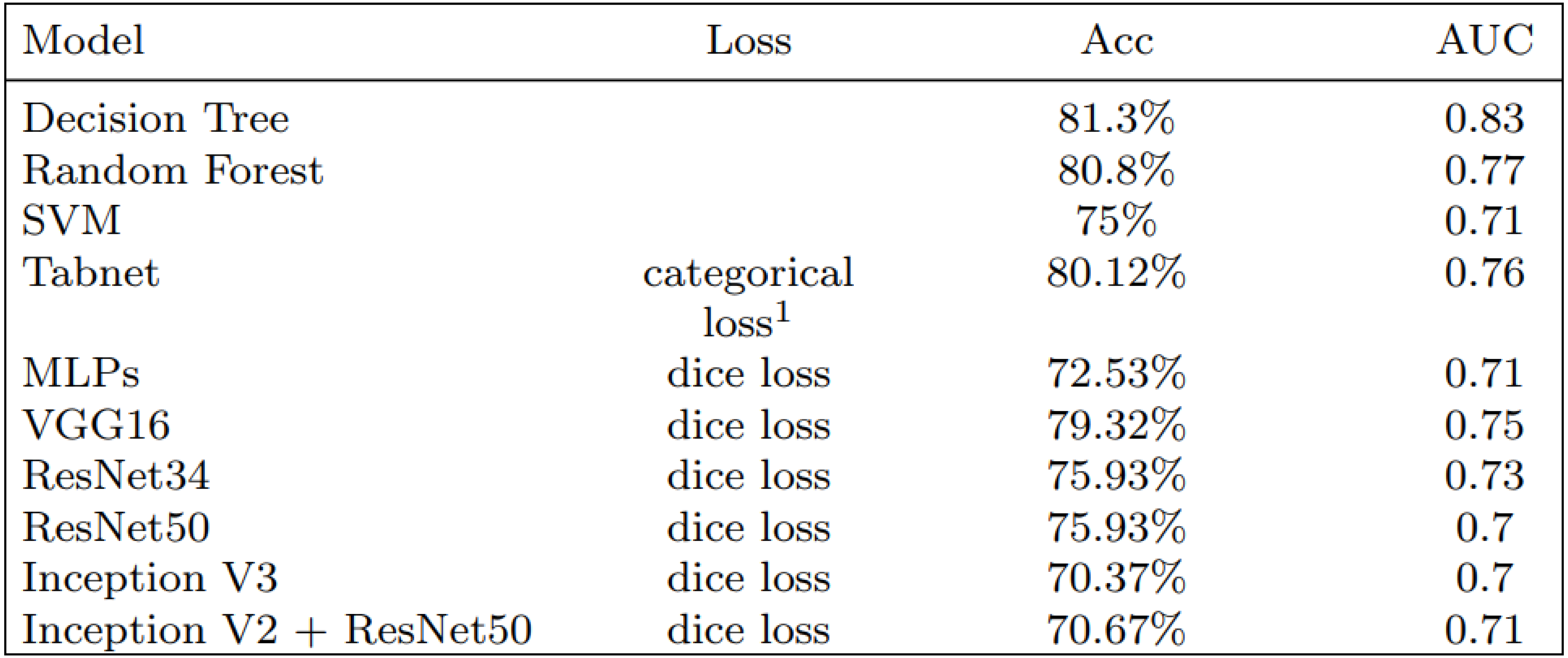


Model	Loss	SMOTE	Acc	AUC
VGG16			72.22%	0.68
VGG16		x	73.77%	0.69
VGG16	focal loss	x	70.06%	0.69
VGG16	f1 loss	x	77.16%	0.71
VGG16	dice loss	x	79.32%	0.75
VGG16	multi loss[1]	x	74.69%	0.73
VGG16	multi loss[2]	x	75%	0.61

[1] F1, dice losses. [2] F1, Focal and Dice losses.

5. EXPERIMENTS AND RESULTS

5.5. Evaluation metrics of all experiment models



Model	Loss	Acc	AUC
Decision Tree		81.3%	0.83
Random Forest		80.8%	0.77
SVM		75%	0.71
Tabnet	categorical loss[1]	80.12%	0.76
MLPs	dice loss	72.53%	0.71
VGG16	dice loss	79.32%	0.75
ResNet34	dice loss	75.93%	0.73
ResNet50	dice loss	75.93%	0.7
Inception V3	dice loss	70.37%	0.7
Inception V2 + ResNet50	dice loss	70.67%	0.71

6. CONCLUSION

- **DL models show potential to overcome limitations of traditional ML**
- **Stable, accurate results through data sampling, modern loss functions**
- **CNN models feasible for fraud prediction with proper techniques**
- **Proves viability of DL in this domain**

Future Work:

- Explore advanced DL techniques to further improve performance
- Expand range of datasets to increase objectivity of research

THANK YOU