Journal of Advanced Research in Fluid Mechanics and Thermal Sciences

Journal homepage:
https://semarakilmu.com.my/journals/index.php/fluid_mechanics_thermal_sciences/index
ISSN: 2289-7879

Enhanced Thermal Comfort Prediction Model by Addressing Outliers and Data Imbalance

Hui-Hui Tan[1], Yi-Fei Tan[1,*], Wooi-Haw Tan[1], Chee-Pun Ooi[1]

[1] Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 1 January 2024
Received in revised form 10 May 2024
Accepted 21 May 2024
Available online 15 June 2024

Keywords:
Isolation forest; SMOTE; ASHRAE Comfort Database II; machine learning; thermal comfort prediction

ABSTRACT

The relationship between individuals and their thermal environment is pivotal not only for comfort but also for health and productivity. Thermal comfort, as defined by ASHRAE, reflects an individual's satisfaction with their ambient thermal conditions and can be gauged using the ASHRAE scale. In the past, traditional thermal comfort prediction models such as the Predicted Mean Vote (PMV) were used to evaluate thermal comfort. Nevertheless, the emergence of machine learning provides a more dynamic approach to predict thermal comfort of occupants. However, the subjective nature of thermal comfort introduces data ambiguities challenge which lead to the existence of outliers. Moreover, data imbalances within the dataset can cause the machine learning models to not learn the minority class effectively, resulting in the deterioration of the model. This research has developed an enhanced thermal comfort prediction model to predict the occupant's thermal comfort by leveraging the outlier detection technique and synthetic data generator, particularly the Isolation Forest and SMOTE. The experiment showed that the proposed model is able to achieve an accuracy of 74.94%. This exhibited a slight improvement compared to the findings in prior research of using Random Forest prediction model.

1. Introduction

The necessity to maintain a harmonious interaction between individuals and their thermal environment is not merely a matter of comfort, but a prerequisite for maintaining good health and well-being. Numerous studies have been conducted to assess thermal comfort in buildings, including studies done by Al-Absi *et al.,* [1], Djabir *et al.,* [2], and Alias *et al.,* [3], further emphasizing the importance of thermal comfort. Poor thermal comfort can bring many side effects on occupants, ranging from reduced work performance to health concerns. Tham and Willem [4] demonstrated that thermal discomfort can undermine an occupant's ability to perform tasks effectively. This sentiment is echoed by Wyon [5], who emphasized that unsatisfactory thermal conditions can directly diminish productivity. The impact is not just confined to physical comfort and work efficiency; cognitive functionality also sees a decline in suboptimal conditions, as indicated by Maddalena *et al.,* [6]. From

[*] *Corresponding author.*
E-mail address: yftan@mmu.edu.my

https://doi.org/10.37934/arfmts.118.1.5264

a health perspective, the Centers for Disease Control and Prevention (CDC) [7] cautioned that occupants exposed to poor thermal environments are at an increased risk of ailments, including hypothermia. These findings underline the significance of maintaining optimal thermal conditions for both the well-being and efficiency of occupants. Hence, thermal comfort becomes a vital metric to determine whether individuals find their surroundings comfortable or not from a thermal perspective to maintain good health. In order to measure an individual thermal comfort level, ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) standard is used [8]. ASHRAE defines thermal comfort as an individual's satisfaction with their surrounding thermal conditions. The ASHRAE scale is ranging from -3 to +3, where -3 is labeled as "Cold", -2 as "Cool", -1 as "Slightly Cool", 0 as "Neutral", 1 as "Slightly Warm", 2 as "Warm", and 3 as "Hot". In the process of collecting an individual thermal comfort level, they are giving their feedback based on ASHRAE scale. This feedback referred as the Thermal Sensation Vote (TSV) [9].

Besides using ASHRAE standard to collect an individual thermal comfort, Fanger introduced Predicted Mean Vote (PMV) model in the late 1960s. This model was built from experiments conducted in climate chambers, supplemented by heat balance equations [10,11], which can be used to evaluate thermal comfort. Recently, advancements have seen the integration of machine learning in the realm of thermal comfort prediction. Unlike traditional methods, this technique empowers computers to self-learn without intricate programming [12]. The literature review Section 1.1 will delve into past research that utilized machine learning for thermal comfort modeling. It is noteworthy, however, that leveraging machine learning in this domain presents unique challenges, especially given the individual variations in thermal comfort.

Thermal comfort is notably subjective, with perceptions differing from one individual to another [13]. This subjectivity creates ambiguity in thermal comfort datasets. It is crucial to emphasize that such ambiguity is not due to data collection errors but arises from the inherent personal preferences regarding thermal comfort. This variability poses challenges to machine learning models, as the outliers can occur due to this variability and it may mislead the data analysis results [14]. Moreover, it could also degrade the performance of machine learning algorithms [15,16]. Hence, data cleaning becomes essential to remove the outliers so that the performance of thermal comfort prediction models can be significantly boosted. Another challenge is data imbalance in the thermal comfort dataset. If feedback from occupants is not adequately collected for all TSV categories, the model may not train effectively. As a result, the data imbalance can negatively affect the model's performance [17-19]. Therefore, implementing data balancing, ensuring uniform distribution across all classes, becomes vital to further optimize the model's performance [20]. Hence, in this study, an enhanced thermal comfort prediction model is developed with the application of Isolation Forest (IF) algorithm to remove the outliers and SMOTE to generate and feed in the synthetic data into the dataset to balance the distribution of data. The study starts with an introduction on the relevance of thermal comfort, followed by a literature review detailing the existing relevant works of building the thermal comfort models. The methodology section explains the steps from data collection to model evaluation. The performance of the models and their implications are then discussed in the results and discussion section. Finally, the paper concludes with a summarization of the findings.

1.1 Literature Review

Thermal comfort prediction is an important area of research, and many studies use machine learning and deep learning to improve predictions. As these methods get better, researchers face new challenges. One of the issues is that sometimes, there is uneven data, meaning some thermal comfort conditions have lots of data, while others have very little. This can affect the model's performance. Additionally, it is crucial to have accurate data. If there are unusual or ambiguous values in the data, it can influence the learning of machine learning models. In this section, recent research on outlier detection techniques, imbalanced data handling techniques, machine learning thermal comfort models, deep learning thermal comfort models, thermal comfort models with transfer learning, and studies addressing data imbalance in thermal comfort models are reviewed.

1.1.1 Outlier detection techniques

Outliers or anomalies in data sets can significantly influence the outcome of data analyses and modeling efforts. Recent years have seen a surge in advanced outlier detection techniques, specifically designed to cater to diverse application domains. In 2021, a study by Nascimento *et al.,* [21] undertook the task of identifying outliers in power consumption data for a tertiary building in France. The study made comparisons between traditional statistical methods, like boxplots, applied directly on measurements and on the deviation between measurements and their respective predictions. Another research in 2021 by Yu *et al.,* [22] revolved around the domain of hyperspectral anomaly detection. The primary objective was to enhance the differentiation between anomalous targets and the typical background. To realize this, the research leveraged an adapted version of the Local Outlier Factor (LOF), which refined the selection of dictionary atoms. Additionally, a specifically designed filter emphasized the spatial structure of data, further enhancing the detection capability. In a 2020 study by Mahajan *et al.,* [23], researchers highlighted the need for accurate air quality data to make decisions about air pollution. Given the constant flow of data, there is a growing need for methods to spot outliers in real-time. The study presented a framework that compared five statistical methods to detect unusual data points in ongoing air quality data.

Lastly, delving into cybersecurity, 2021 research by Heigl *et al.,* [24] introduced the Performance Counter Based iForest framework. This innovative framework employed variants of the Isolation Forest (IF) algorithm to detect outliers, specifically focusing on identifying malicious activities in real-world computer networks. The emphasis on using multiple variants of the Isolation Forest showcases the adaptability and efficiency of this method in a real-time detection scenario.

1.1.2 Imbalanced data handling techniques

Researchers have explored various data balancing techniques to mitigate the negative impact of imbalanced datasets on prediction model accuracy. The following examples highlight the techniques utilized across different areas by researchers. Low *et al.,* [25] developed a Commercial Vehicle Activity prediction model employing a gradient boosting approach, enhanced with data resampling techniques including random undersampling, random oversampling, and the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. A study by Karatas *et al.,* [20] focused on developing Intrusion Detection Systems (IDSs) using six machine learning algorithms, employing SMOTE to overcome data imbalance. In a study by Kaya *et al.,* [26] the impact of classification algorithms, feature selection, and data balancing methods, particularly SMOTE-based techniques, on Software Vulnerability Prediction models was investigated. A study by Ivanciu *et al.,* [27] examined three data balancing strategies—SMOTE with Tomek Links, SMOTE with outlier elimination, and random subsampling—in the context of electronic payment transactions, where fraudulent operations are the minority class. Finally, a study by Uttam and Sharma [28] investigated the effectiveness of random oversampling, SMOTE, and SMOTE Tomek in detecting credit card fraud using Neural Networks. As such, it can be concluded that SMOTE is a popular technique to handle imbalance data.

1.1.3 Machine learning thermal comfort models

The study and optimization of thermal comfort has become an area of keen interest in recent years, and various research efforts have employed machine learning techniques to better predict and understand comfort levels based on various environmental and individual factors. In 2017, a study by Chaudhuri *et al.,* [29] employed ASHRAE RP-884 dataset focusing solely on data from Singapore to develop their model. The model used a comprehensive set of inputs, including Air Temperature, Mean Radiant Temperature, and several individual factors such as Gender and Age. The researchers employed various algorithms, notably Machine Learning Classifiers, PMV model, Extended PMV, and Adaptive PMV. The highest accuracy achieved among these was 85.3% with the Machine Learning Classifiers, while the Adaptive PMV model demonstrated a notably lower efficiency of 35.51%.

A subsequent study in 2021 by Abdulgader and Lashhab [30], also utilizing the ASHRAE RP-884 dataset, incorporated parameters like Outdoor Temperature and Standard Effective Temperature to predict Thermal Comfort Value. They employed a diverse set of algorithms, including Multiple Liner Regression (MLR), Support Vector Regression (SVR), Random Forest Regression (RFR), and Decision Tree Regression (DTR). Among these, the SVR algorithm achieved the best results, with Root Mean Square Error (RMSE) value of 0.7601 and an R2 Score of 0.3766. In 2022, a different approach was seen in the study by Acquaah *et al.,* [31] where the researchers used their own dataset. This study stood out for its multiple labels with multiple classes output, including thermal comfort and thermal sensation, among others. The employed algorithms – Extratrees, Random Forest, Decision trees, and K-nearest neighbour – achieved varying degrees of accuracy. Extratrees demonstrated the best performance with an accuracy of 68% and an MSE of 2.15. The year 2023 saw further advancements. In a study by Feng *et al.,* [32], a hybrid ensemble learning approach was utilized. This research uniquely combined two or more algorithms, leveraging the strengths of Extreme Learning Machine (ELM), Stochastic configuration network (SCN), Random Forest (RF), Support vector regression (SVR) to achieve impressive RMSE values ranging from 0.157 to 0.237. Another 2023 study by Tekler *et al.,* [33] presented results from their own dataset, predicting user preferences for room temperature adjustments using the Extreme Gradient Boosting (XGB) algorithm. The output classes were distinctly labeled as 'Cooler', 'No Change', and 'Warmer', with an achieved accuracy of around 75%.

1.1.4 Deep learning thermal comfort models

Deep learning methods, a subset of machine learning techniques that utilize multi-layered neural networks, have garnered significant attention in predicting thermal comfort based on various environmental and individual parameters. In 2019, a study by Ma *et al.,* [34] utilized the ASHRAE RP-884 dataset to forecast the Thermal Comfort Value. This research stood out for its inclusion of parameters like Distance Between People and Equipment, Human Activity Type, and individual characteristics such as Gender, Age, Weight, and Height. Using Artificial Neural Network (ANN) and the PMV Model, they achieved Mean Square Error (MSE) values of 0.39 and 2.1 respectively, showing the superiority of ANN in this particular context. The subsequent year, 2020, Irshad *et al.,* [35] developed their model using their own dataset, which considered parameters like globe temperature and clothing value. Using ANN, they achieved a compellingly low MSE of 0.07956. In contrast, Gao *et al.,* [36], using the ASHRAE RP-884 dataset, employed algorithms including Deep Feedforward Neural Network, SVM, and others. Among these algorithms, the Deep Feedforward Neural Network recorded the lowest MSE value of 1.1583. A unique perspective was brought to the field in year 2021. In a study by Brik *et al.,* [37], the focus was shifted towards understanding the indoor thermal comfort of disabled individuals. This research incorporated parameters like Disability Type and several indoor environmental factors. Among the algorithms used, the Artificial Neural Classifier (ANC) stood out with an astounding accuracy of 94%, significantly outperforming traditional classifiers like Logistic Regression and Decision Tree. Lastly, in 2022, a study by Lala *et al.,* [38] utilized both the ASHRAE Global Thermal Comfort Database II and their own dataset, focused on primary school students. Their proposed model, DeepComfort, was compared with several single-task techniques, and it excelled particularly concerning F1-score, Precision, and Recall. The emphasis on F1-score as a performance metric was due to existing data imbalances, which can mask the accuracy of predicting minority classes. Remarkably, DeepComfort outperformed even the Bayesian deep neural network, further underscoring its effectiveness.

1.1.5 Thermal comfort models with transfer learning

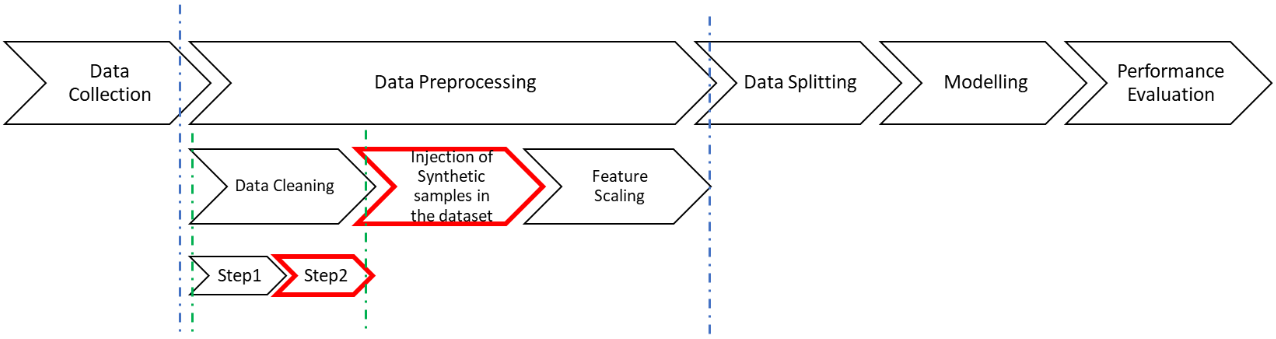
There are researchers who explore the development of thermal comfort prediction models through the application of transfer learning approaches. In 2021, a study by Gao *et al.,* [39] emphasized the utility of transfer learning, a technique that apply knowledge gained from one domain to enhance performance in a related domain. When comparing the Transfer Learning Multilayer Perception (TL-MLP) to traditional models like Random Forest and SVM, it achieved a commendable accuracy range of 50.76 – 54.50%, underscoring the potential of transfer learning in this field. Transfer learning was again the focal point of research by Somu *et al.,* [40] in the same year. They have selected the ASHRAE RP-884 and Scales Project datasets as the source domains, and the Medium US Office dataset as the target dataset. Among numerous deep learning models, the Transfer Learning CNN-LSTM model outshined others with an accuracy of 59.84%. In 2022, another study by Park and Park [41] proposed an ensemble transfer learning (TL) approach for their thermal comfort prediction model, aiming to transfer knowledge across datasets from different indoor spaces and thermal environments. This study utilized a dataset collected by the researchers themselves. The results demonstrated that the ensemble TL approach enhanced the accuracy of thermal comfort predictions for two target subjects using a model pre-trained on a source dataset. In 2023, Zhang and Li [42] proposed integrating transfer learning with a transformer model to predict thermal comfort, utilizing the ASHRAE RP-884 dataset from the Scales project as the source and the Medium US dataset as the target domain. The proposed TL-Transformer model achieved an accuracy of 62.6%, outperforming other state-of-the-art methods tested in their experiments. Moreover, Natarajan and Laftchiev [43] introduced a transfer active learning framework for thermal comfort prediction, significantly reducing the need for a large, labeled dataset. The study, conducted with a dataset collected by the authors, demonstrated that their methodology achieved a mean error of approximately 0.82, which is lower than that of traditional supervised learning models.

1.1.6 Studies addressed data imbalance in thermal comfort models

There are also researchers who explored building thermal comfort prediction model by addressing the imbalanced data issue. Study by Fayyaz *et al.,* [44] in 2021 dealt with the data imbalance by deploying both downsampling and oversampling techniques. By focusing on HVAC building data from the ASHRAE RP-884 dataset, the researchers found that SVM and RF, after addressing the imbalance, could achieve accuracies up to 86.08%, marking a significant improvement. In 2022, a study by Cakir and Akbulut [45] that employed the ASHRAE Comfort Database II, SMOTE was used to address data imbalance. Their results highlighted that Deep Neural Network (DNN) achieved the highest accuracy of 78.01%, outperforming traditional models like Gradient Boosting and PMV. Finally, a study by Martins *et al.,* [46] in 2022 emphasized the role of downsampling to address data imbalance. Their model, which considered parameters like health perception, showed that a DNN with health perception could achieve an accuracy of up to 91.67%. This suggests that considering additional human-centric parameters, when combined with techniques to address data imbalance, can offer a more comprehensive and accurate model for thermal comfort prediction.

2. Methodology

In this section, an overview of the systematic procedure for developing a reliable thermal comfort prediction model is presented. As depicted in Figure 1, this procedure consists of 5 stages. Firstly, the data acquisition is conducted, followed by data preprocessing, data splitting, model development, and performance evaluation. Within the data preprocessing process, there are 3 sub-processes involved, which include data cleaning, injection of synthetic samples in the dataset, and feature scaling. The data cleaning process is further divided into 2 steps, which are identifying and eliminating unimportant or missing data columns (Step 1) and outlier detection (Step 2). Later, the data imbalance issue is addressed by injection of synthetic samples in the dataset processes. Lastly, feature scaling is implemented. A detailed explanation of each of these stages is provided in subsections 2.1 through 2.5.



Step 1: Identifying and eliminating unimportant or missing data columns
Step 2: Outlier Detection

Fig. 1. Thermal comfort prediction framework

2.1 Data Acquisition

In this study, the ASHRAE dataset obtained from Cakir and Akbulut [45] is the selected dataset for building the thermal comfort prediction model. This dataset consists of dataset consists of 40,988 rows and 17 features. The features are Clothing Insulation, Metabolic Rate, Standard Effective Temperature, Air Temperature, Globe Temperature, Air Velocity, Relative Humidity, Outdoor Monthly Air Temperature, Publication (Citation), Year, Season, PMV, Koppen climate classification, Building type, Cooling strategy, building level, Sex and Age.

2.2 Data Preprocessing

The purpose of data preprocessing is to clean, scale and ensure the data is ready for the model learning. It consists of 3 phases which include Data Cleaning, Handling Imbalanced Data, and Feature Scaling.

2.2.1 Data cleaning

This process begins by identifying and eliminating unimportant or missing data columns. Following that, an outlier technique will be applied to remove unusual data points. The detailed explanations are as below:

Step 1: Identifying and eliminating unimportant or missing data columns

In accordance with the input features suggested by Cakir and Akbulut [45], the development of a predictive model for thermal comfort requires important features from personal factors, indoor environment factors, and outdoor environment factors. Those features include Clothing Insulation, Metabolic Rate, Standard Effective Temperature, Air Temperature, Globe Temperature, Air Velocity, Relative Humidity, and Outdoor Monthly Air Temperature. The target variable chosen for this predictive model is Thermal Sensation with 7 classes, ranging from cold to hot, which is known as “7-point TSV”. While all other columns, are removed from the dataset. The cleaned dataset consists of 9 columns with 8 input features and 1 output.

Step 2: Outlier Detection

Among the outlier detection techniques, IF is chosen for this study [47]. One of the reasons of choosing IF is it is a highly efficient algorithm especially for large dataset. Furthermore, there is a parameter called "contamination" in IF that users can modify to decide the ratio of outliers in the dataset to be removed. This parameter allows the value range (0, 0.5]. Different contamination values, 0.15, 0.25, 0.35, and 0.45 will be experimented with to investigate which one works best for building the thermal comfort prediction model.

Following the application of the IF with various contamination values, each contamination value provides a different total number of data points. The specific quantities of data points corresponding to each contamination value are presented in Table 1.

Table 1
Number of data points for each contamination value

Contamination	Number of Data Points
0.15	34,839
0.25	30,741
0.35	26,642
0.45	22,543

2.2.2 Injection of synthetic samples in the dataset

The dataset, both before and after the applying IF algorithm, exhibits significant imbalance. To mitigate this imbalance and enhance the reliability of the proposed model, SMOTE is employed to generate synthetic samples, thereby increasing the amount of data in the dataset and achieving a more balanced distribution. The amount of data before and after applying SMOTE across various contamination values and without applying Isolation Forest are visualized in Figure 2 and Figure 3.

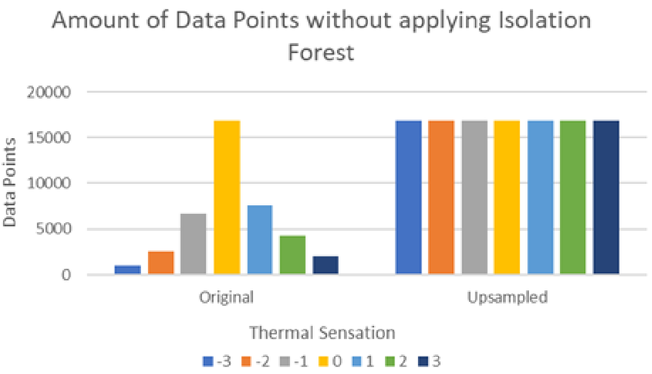
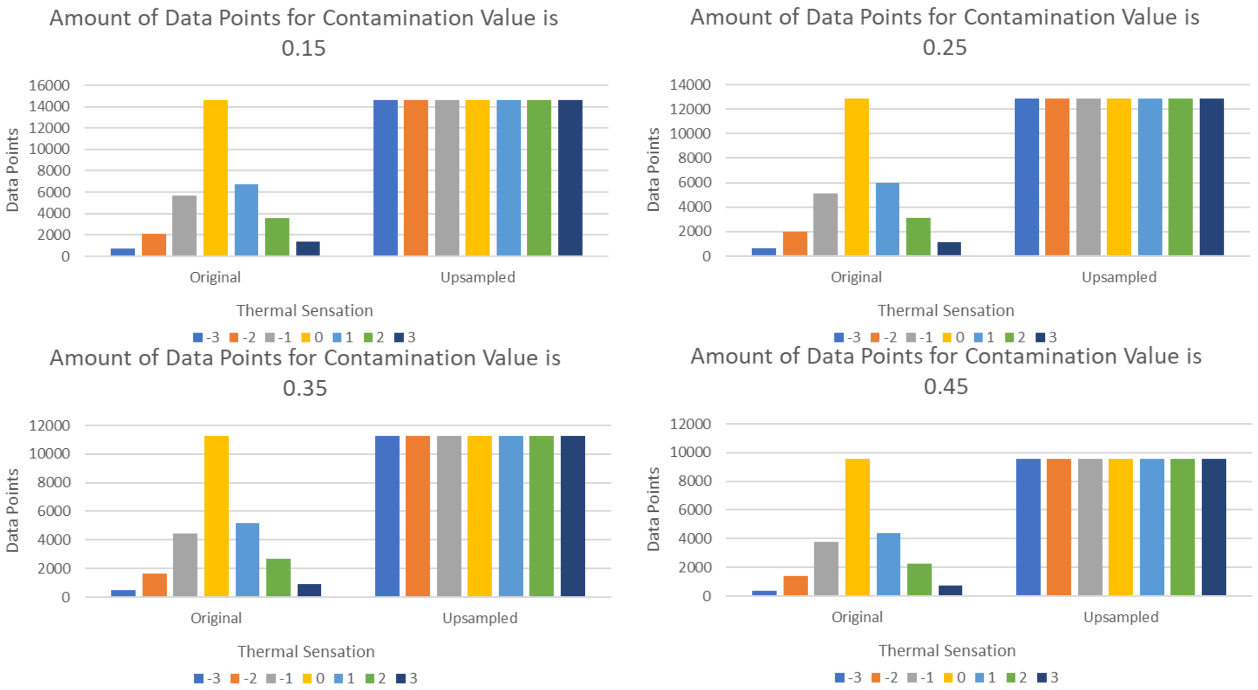

Amount of Data Points without applying Isolation Forest

Fig. 2. Amount of Data Points without applying IF


Amount of Data Points for Contamination Value is 0.15

Amount of Data Points for Contamination Value is 0.25

Amount of Data Points for Contamination Value is 0.35

Amount of Data Points for Contamination Value is 0.45

Fig. 3. Amount of Data Points by applying IF with Contamination Values of 0.15, 0.25, 0.35, and 0.45

2.2.3 Feature scaling

Feature scaling serves as a procedure in the standardization of independent variables within a dataset. This operation is important in maintaining equal distribution of influence from each variable during the model's training process. For this study, we utilized the Standard Scaler, outlined in (1), to fit and transform our training dataset. Identical Standard Scaler was then applied to the test dataset, ensuring a uniform scaling application across both sets of data. This meticulous procedure was adopted to prevent any potential data leakage.

$$x_{stand} = \frac{x - mean(x)}{standard\ deviation(x)} \tag{1}$$

2.3 Data Splitting

For effective model evaluation, the dataset will be divided into two segments: 90% for training and 10% for testing, utilizing K-Fold validation with K set to 10, this is to ensure a robust assessment of our model's generalization capability. In pursuit of maintaining a consistent data distribution across these cross-validation folds, Stratified K-Fold is selected to preserve the proportional representation of different classes in the dataset, thereby minimizing variance in the model evaluation.

2.4 Model Development

In the development of thermal comfort prediction models, two types of models are used, namely tree-based ensemble model, and distance-based model. The particular classifiers chosen for the two types of models are Random Forest, and K-Nearest Neighbors respectively. Notably, all the models used in this study were configured with default settings from the Scikit-Learn library for thermal comfort prediction model development.

2.5 Performance Evaluation

The effectiveness of the model will be assessed by using key performance indicators such as accuracy, precision, recall, and the F1 score. Accuracy provides the proportion of correctly classified instances out of the total instances in a dataset. Whereas, precision and recall measure the proportion of true positive predictions (correctly identified positive cases) out of all positive predictions made and the proportion of true positive predictions out of all actual positive cases, respectively. Finally, the F1 Score is a balance between precision and recall, providing a single metric that combines both.

3. Results and Discussions

This section provides an analysis of the performance of 7-point TSV prediction models using different types of machine learning algorithms based on various contamination values in Isolation Forest. Figure 4 and Figure 5 display the performance of these models. It is important to note that the results for each machine learning algorithm were obtained after training on the same dataset using 10-fold cross-validation.

From the figures, Random Forest exhibits a mean accuracy that spans from 74.28% to 74.94%. This suggests a consistent performance across different contamination values. The K-Nearest Neighbor model displays a mean accuracy range of 62.58% to 63.62%. This is notably lower than the Random Forest model by approximately 11 percentage points. Precision for the Random Forest model varies between 73.57% and 74.30%, indicating a relatively stable output across the contamination values. The K-Nearest Neighbor model has a precision ranging from 60.74% to 62.14%, again trailing the Random Forest by over 11 percentage points. The Random Forest model's recall ranges between 74.28% and 74.94%, aligning closely with its accuracy. The recall for the K-Nearest Neighbor model oscillates between 62.58% and 63.62%, mirroring its accuracy values. The F1 score, representing the harmonic mean of precision and recall, for the Random Forest model fluctuates between 73.81% and 74.52%. For the K-Nearest Neighbor model, the F1 score ranges from 60.95% to 62.33%. Throughout the metrics of accuracy, precision, recall, and F1 score, the Random Forest model consistently surpasses the K-Nearest Neighbor model. The performance metrics for the Random Forest model suggest a stable response to changes in contamination values with narrow ranges in all metrics. In contrast, the K-Nearest Neighbor model, while presenting some degree of consistency, yields lower performance values across all metrics. The Random Forest with Isolation Forest contamination value of 0.15 achieves the best performance among all the models, which slightly surpasses the findings obtained in the study by Cakir and Akbulut [45] of their Random Forest model.

As the contamination value in the Isolation Forest increases for the Random Forest algorithm, we observe a unique performance trend: the model's performance first improves, then diminishes, and subsequently improves again. However, with the K-Nearest Neighbor algorithm, the performance initially drops with increased contamination value but eventually rebounds. The general trend across all models indicates that, post-implementation of the Isolation Forest, there is an initial decrease in performance, which is then followed by an improvement. This phenomenon suggests that at lower contamination values, the Isolation Forest might be discarding some pertinent data. However, as we increase the contamination value, model performance improves due to more selective data. Users can select a contamination value that achieves optimal model performance while retaining a data volume appropriate for their research.

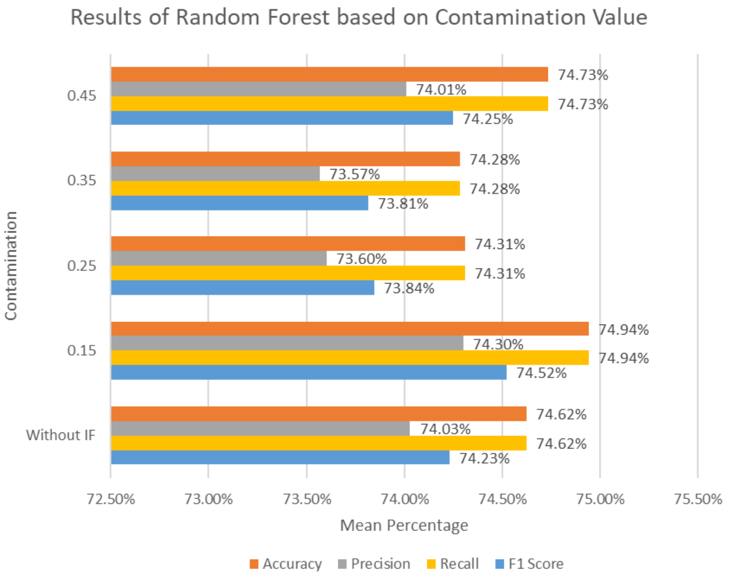


Fig. 4. Results of Random Forest based on Contamination Value

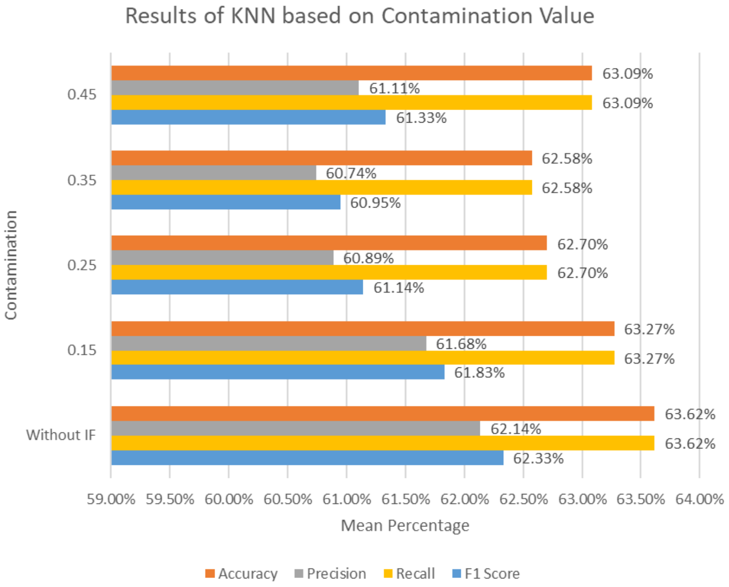


Fig. 5. Results of KNN based on Contamination Value

The standard deviation for all metrics within the Random Forest models ranges between 0.003785574 and 0.00527, whereas for the K-Nearest Neighbor models, it lies between 0.003207 and 0.004547. This uniformity implies an absence of variance during model development.

4. Conclusions

This study demonstrates that the enhanced thermal comfort prediction model that integrating outlier detection with the isolation forest and the injection of synthetic samples in the dataset using SMOTE in the data preprocessing slightly outperformed the results from the study by Cakir and Akbulut [45]. The performance trend, based on the contamination value in the Isolation Forest, suggests a risk in the original data where critical data might be discarded. However, by adjusting the contamination value in the isolation forest, researchers can achieve optimal model performance and preserve essential data. The small standard deviation across all metrics for every model indicates an absence of variance during model development.

In future research, potential areas of improvement include addressing any class overlapping within the thermal comfort dataset, using deep learning models to learn better from complex datasets, and conducting hyperparameter tuning to further optimize model performance.

Acknowledgement

This research is financially supported by TM R&D, Malaysia (RDTC/221057).

References

[1] Al-Absi, Zeyad Amin, Mohd Isa Mohd Hafizal, and Mazran Ismail. "Field Assessment of The Indoor Thermal Environment and Thermal Comfort Levels for Naturally Conditioned Residential Buildings in The Tropical Climate." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 100, no. 3 (2022): 51-66. https://doi.org/10.37934/arfmts.100.3.5166

[2] Djabir, Djabir Abdoulaye, Azian Hariri, Mohamad Nur Hidayat Mat, and Md Hasanuzzaman. "Thermal comfort of indoor open spaces at university library in Malaysia." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 94, no. 2 (2022): 142-165. https://doi.org/10.37934/arfmts.94.2.142165

[3] Alias, Noorazimah Mat, Umar Kassim, Sinar Arzuria Adnan, Nur Hidayah Ahmad Zaidi, Siti Nurul Aqmariah Mohd Kanafiah, Mohd Nazaruddin Yusoff, and Sk Muiz Sk Abd Razak. "Analysis of Thermal Comfort Among Workshop Users: At TVET Technical Institution." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 114, no. 2 (2024): 205-213. https://doi.org/10.37934/arfmts.114.2.205213

[4] Tham, Kwok Wai, and Henry Cahyadi Willem. "Room air temperature affects occupants' physiology, perceptions and mental alertness." *Building and Environment* 45, no. 1 (2010): 40-44. https://doi.org/10.1016/j.buildenv.2009.04.002

[5] Wyon, David P. "The effects of indoor air quality on performance and productivity." *Indoor Air* 14 (2004). https://doi.org/10.1111/j.1600-0668.2004.00278.x

[6] Maddalena, R., M. J. Mendell, K. Eliseeva, W. R. Chan, D. P. Sullivan, M. Russell, U. Satish, and W. J. Fisk. "Effects of ventilation rate per person and per floor area on perceived air quality, sick building syndrome symptoms, and decision-making." *Indoor Air* 25, no. 4 (2015): 362-370. https://doi.org/10.1111/ina.12149

[7] Centers for Disease Control and Prevention (CDC). "Hypothermia-related deaths--Philadelphia, 2001, and United States, 1999." *MMWR. Morbidity and Mortality Weekly Report* 52, no. 5 (2003): 86-87.

[8] ASHRAE. "Thermal environmental conditions for human occupancy." *ANSI/ASHRAE, 55 5* (1992).

[9] Kohoutková, Alžběta, Jana Horváthová, Martin Kny, and Ondřej Nehasil. "Case study of occupant's perception of indoor thermal conditions under different heating systems." In *2018 Building Performance Analysis Conference and SimBuild*, pp. 205-212. 2018.

[10] Yang, Liu, Haiyan Yan, and Joseph C. Lam. "Thermal comfort and building energy consumption implications-a review." *Applied Energy* 115 (2014): 164-173. https://doi.org/10.1016/j.apenergy.2013.10.062

[11] Song, Ying, Fubing Mao, and Qing Liu. "Human comfort in indoor environment: a review on assessment criteria, data collection and data analysis methods." *IEEE Access* 7 (2019): 119774-119786. https://doi.org/10.1109/ACCESS.2019.2937320

[12] Brown, Sara. "Machine learning, explained." *MIT Sloan School of Management*. April 21, 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[13] Fountain, Marc, Gail Brager, and Richard De Dear. "Expectations of indoor climate control." *Energy and Buildings* 24, no. 3 (1996): 179-182. https://doi.org/10.1016/S0378-7788(96)00988-7

[14] Zhu, Jinlin, Zhiqiang Ge, Zhihuan Song, and Furong Gao. "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data." *Annual Reviews in Control* 46 (2018): 107-133. https://doi.org/10.1016/j.arcontrol.2018.09.003

[15] McClelland, Gary H. *Nasty data: Unruly, ill-mannered observations can ruin your analysis*. Cambridge University Press, 2014. https://doi.org/10.1017/CBO9780511996481.028

[16] Frénay, Benoît, and Michel Verleysen. "Reinforced extreme learning machines for fast robust regression in the presence of outliers." *IEEE Transactions on Cybernetics* 46, no. 12 (2015): 3351-3363. https://doi.org/10.1109/TCYB.2015.2504404

[17] Anwar, Muhammad Nafees. "Complexity measurement for dealing with class imbalance problems in classification modelling." *PhD diss., Massey University, New Zealand*, 2012.

[18] Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks* 106 (2018): 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

[19] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357. https://doi.org/10.1613/jair.953

[20] Karatas, Gozde, Onder Demir, and Ozgur Koray Sahingoz. "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset." *IEEE Access* 8 (2020): 32150-32162. https://doi.org/10.1109/ACCESS.2020.2973219

[21] Nascimento, Gustavo Felipe Martin, Frédéric Wurtz, Patrick Kuo-Peng, Benoit Delinchant, and Nelson Jhoe Batistela. "Outlier Detection in Buildings' Power Consumption Data Using Forecast Error." *Energies* 14, no. 24 (2021): 8325. https://doi.org/10.3390/en14248325

[22] Yu, Shaoqi, Xiaorun Li, Liaoying Zhao, and Jing Wang. "Hyperspectral anomaly detection based on low-rank representation using local outlier factor." *IEEE Geoscience and Remote Sensing Letters* 18, no. 7 (2020): 1279-1283. https://doi.org/10.1109/LGRS.2020.2994745

[23] Mahajan, Manish, Santosh Kumar, Bhasker Pant, and Umesh Kumar Tiwari. "Incremental outlier detection in air quality data using statistical methods." In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1-5. IEEE, 2020. https://doi.org/10.1109/ICDABI51230.2020.9325683

[24] Heigl, Michael, Kumar Ashutosh Anand, Andreas Urmann, Dalibor Fiala, Martin Schramm, and Robert Hable. "On the improvement of the isolation forest algorithm for outlier detection with streaming data." *Electronics* 10, no. 13 (2021): 1534. https://doi.org/10.3390/electronics10131534

[25] Low, Raymond, Lynette Cheah, and Linlin You. "Commercial vehicle activity prediction with imbalanced class distribution using a hybrid sampling and gradient boosting approach." *IEEE Transactions on Intelligent Transportation Systems* 22, no. 3 (2020): 1401-1410. https://doi.org/10.1109/TITS.2020.2970229

[26] Kaya, Aydin, Ali Seydi Keceli, Cagatay Catal, and Bedir Tekinerdogan. "The impact of feature types, classifiers, and data balancing techniques on software vulnerability prediction models." *Journal of Software: Evolution and Process* 31, no. 9 (2019): e2164. https://doi.org/10.1002/smr.2164

[27] Ivanciu, Laura-Nicoleta, Adelina-Veronica Dumitras, and Emilia Sipos. "Analysis of Data Balancing Techniques in Fraudulent Transactions Datasets." In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1-4. IEEE, 2023. https://doi.org/10.1109/ISMSIT58785.2023.10304932

[28] Uttam, Atul Kumar, and Gaurav Sharma. "A comparison of data balancing techniques for credit card fraud detection using neural network." In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 1136-1140. IEEE, 2021. https://doi.org/10.1109/I-SMAC52330.2021.9640911

[29] Chaudhuri, Tanaya, Yeng Chai Soh, Hua Li, and Lihua Xie. "Machine learning based prediction of thermal comfort in buildings of equatorial Singapore." In *2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC)*, pp. 72-77. IEEE, 2017. https://doi.org/10.1109/ICSGSC.2017.8038552

[30] Abdulgader, Musbah, and Fadel Lashhab. "Energy-efficient thermal comfort control in smart buildings." In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0022-0026. IEEE, 2021. https://doi.org/10.1109/CCWC51732.2021.9376175

[31] Acquaah, Yaa T., Balakrishna Gokaraju, Raymond C. Tesiero III, and Kaushik Roy. "Machine learning techniques to predict real time thermal comfort, preference, acceptability, and sensation for automation of HVAC temperature."

In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 659-665. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-08530-7_55

[32] Feng, Xue, Eryan Bin Zainudin, Hong Wen Wong, and King Jet Tseng. "A hybrid ensemble learning approach for indoor thermal comfort predictions utilizing the ASHRAE RP-884 database." *Energy and Buildings* 290 (2023): 113083. https://doi.org/10.1016/j.enbuild.2023.113083

[33] Tekler, Zeynep Duygu, Yue Lei, Yuzhen Peng, Clayton Miller, and Adrian Chong. "A hybrid active learning framework for personal thermal comfort models." *Building and Environment* 234 (2023): 110148. https://doi.org/10.1016/j.buildenv.2023.110148

[34] Ma, Guofeng, Ying Liu, and Shanshan Shang. "A building information model (BIM) and artificial neural network (ANN) based system for personal thermal comfort evaluation and energy efficient design of interior space." *Sustainability* 11, no. 18 (2019): 4972. https://doi.org/10.3390/su11184972

[35] Irshad, Kashif, Asif Irshad Khan, Sayed Ameenuddin Irfan, Md Mottahir Alam, Abdulmohsen Almalawi, and Md Hasan Zahir. "Utilizing artificial neural network for prediction of occupants thermal comfort: A case study of a test room fitted with a thermoelectric air-conditioning system." *IEEE Access* 8 (2020): 99709-99728. https://doi.org/10.1109/ACCESS.2020.2985036

[36] Gao, Guanyu, Jie Li, and Yonggang Wen. "DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning." *IEEE Internet of Things Journal* 7, no. 9 (2020): 8472-8484. https://doi.org/10.1109/JIOT.2020.2992117

[37] Brik, Bouziane, Moez Esseghir, Leila Merghem-Boulahia, and Hichem Snoussi. "An IoT-based deep learning approach to analyse indoor thermal comfort of disabled people." *Building and Environment* 203 (2021): 108056. https://doi.org/10.1016/j.buildenv.2021.108056

[38] Lala, Betty, Hamada Rizk, Srikant Manas Kala, and Aya Hagishima. "Multi-task learning for concurrent prediction of thermal comfort, sensation and preference in winters." *Buildings* 12, no. 6 (2022): 750. https://doi.org/10.3390/buildings12060750

[39] Gao, Nan, Wei Shao, Mohammad Saiedur Rahaman, Jun Zhai, Klaus David, and Flora D. Salim. "Transfer learning for thermal comfort prediction in multiple cities." *Building and Environment* 195 (2021): 107725. https://doi.org/10.1016/j.buildenv.2021.107725

[40] Somu, Nivethitha, Anirudh Sriram, Anupama Kowli, and Krithi Ramamritham. "A hybrid deep transfer learning strategy for thermal comfort prediction in buildings." *Building and Environment* 204 (2021): 108133. https://doi.org/10.1016/j.buildenv.2021.108133

[41] Park, Hansaem, and Dong Yoon Park. "Prediction of individual thermal comfort based on ensemble transfer learning method using wearable and environmental sensors." *Building and Environment* 207 (2022): 108492. https://doi.org/10.1016/j.buildenv.2021.108492

[42] Zhang, Xin, and Peng Li. "Transfer learning in the transformer model for thermal comfort prediction: a case of limited data." *Energies* 16, no. 20 (2023): 7137. https://doi.org/10.3390/en16207137

[43] Natarajan, Annamalai, and Emil Laftchiev. "A transfer active learning framework to predict thermal comfort." *International Journal of Prognostics and Health Management* 10, no. 3 (2019). https://doi.org/10.36001/ijphm.2019.v10i3.2629

[44] Fayyaz, Muhammad, Asma Ahmad Farhan, and Abdul Rehman Javed. "Thermal comfort model for HVAC buildings using machine learning." *Arabian Journal for Science and Engineering* (2022): 1-16.

[45] Cakir, Mustafa, and Akhan Akbulut. "A bayesian deep neural network approach to seven-point thermal sensation perception." *IEEE Access* 10 (2022): 5193-5206. https://doi.org/10.1109/ACCESS.2022.3140951

[46] Martins, Larissa Arakawa, Veronica Soebarto, Terence Williamson, and Dino Pisaniello. "Personal thermal comfort models: A deep learning approach for predicting older people's thermal preference." *Smart and Sustainable Built Environment* 11, no. 2 (2022): 245-270. https://doi.org/10.1108/SASBE-08-2021-0144

[47] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413-422. IEEE, 2008. https://doi.org/10.1109/ICDM.2008.17