AI

MDPI

Article

Probabilistic Ensemble Framework for Injury Narrative Classification

Srushti Vichare, Gaurav Nanda and Raji Sundararajan *

School of Engineering Technology, Purdue University, West Lafayette, IN 47907, USA; svichar@purdue.edu (S.V.); gnanda@purdue.edu (G.N.)
* Correspondence: raji@purdue.edu

Abstract: In this research, we analyzed narratives from the National Electronic Injury Surveillance System (NEISS) dataset to predict the top two injury codes using a comparative study of ensemble machine learning (ML) models. Four ensemble models were evaluated: Random Forest (RF) combined with Logistic Regression (LR), K-Nearest Neighbor (KNN) paired with RF, LR combined with KNN, and a model integrating LR, RF, and KNN, all utilizing a probabilistic likelihood-based approach to improve decision-making across different classifiers. The combined KNN + LR ensemble achieved an accuracy of 90.47% for the top one prediction, while the KNN + RF + LR model excelled in predicting the top two injury codes with a very high accuracy of 99.50%. These results demonstrate the significant potential of ensemble models to enhance unstructured narrative classification accuracy, particularly in addressing underrepresented cases, and the potential of the proposed probabilistic ensemble framework ML models in improving decision-making in public health and safety, providing a foundation for future research in automated clinical narrative classification and predictive modeling, especially in scenarios with imbalanced data.

Keywords: injury classification; injury narratives; machine learning models; ensemble models

Citation: Vichare, S.; Nanda, G.; Sundararajan, R. Probabilistic Ensemble Framework for Injury Narrative Classification. *AI* **2024**, *5*, 1684–1694. https://doi.org/10.3390/ai5030082

Academic Editors: Daniele Giansanti and Giovanni Costantini

Received: 2 August 2024
Revised: 13 September 2024
Accepted: 14 September 2024
Published: 20 September 2024

1. Introduction

Injury classification categorizes injuries by location, severity, cause, and type, which is an essential part of tracking public health and safety by understanding the underlying causes and circumstances of injuries and, thus, preventing them and minimizing harm. It also helps in achieving better diagnosis and treatment of injuries [1]. Multiple public health agencies and organizations are involved in the collection and classification of injury narratives from hospital emergency rooms (ERs). The narratives collected at ERs are analyzed by these agencies by assigning different types of diagnostic injury codes, such as their nature, cause, and severity. In the US, the Consumer Product Safety Commission (CPSC) maintains the National Electronic Injury Surveillance System (NEISS), which collects comprehensive narratives outlining the circumstances leading to accidents caused by consumer products [2]. The narratives collected at ERs are crucial for understanding causes and concerns with product safety. The narratives allow public health experts to study and track injury patterns, spot new hazards, and create focused preventive measures. After the narratives are categorized into a structured form by assigning injury diagnostic codes, the statistical analysis of classified injury narratives provides insightful information about the common causes and conditions that lead to injuries. However, the manual approach to categorizing these narratives into injury codes is time-consuming and prone to errors. Previous studies have used machine learning (ML)-based approaches for classifying injury narratives, but the accuracy has been observed to be limited for individual ML models due to the large number of prediction categories and noisy nature of narratives [3–7].

This motivated us to study the prediction performance of three state-of-the-art ML models and different ensemble models, based on their probabilistic likelihood of correctness, in classifying product-related injury narratives.

2. Literature Review

Researchers studied the use of narrative analysis for describing the mechanisms of injury and comparing patterns of work-related fatalities in New Zealand, Australia, and the United States [3]. The study underscored the potential of narrative analysis to bridge these gaps by offering a consistent framework for comparing work-related fatalities across countries. The research revealed the narrative analysis's need in examining work-related deaths by offering an innovative perspective alongside conventional coding methods. The analysis improved the knowledge of job safety and aids in creating more focused preventive measures for different countries and sectors. The narrative analysis minimized misclassification by enabling researchers to use narrators' own words, fixing coding inaccuracies in complex cases [4]. While narrative analysis has its merits, it also presents obstacles that necessitate a structured coding process. Thus, ML techniques are seen to be a substitute for the tedious, manual coding of narratives, offering a faster and efficient classification process.

Fuzzy Bayesian models and automated techniques demonstrated promising results in the past, precisely classifying cause-of-injury codes from narratives and in the capacity to separate out cases for human evaluation [5,6]. Classifying injury narratives from large administrative databases has been made easier with a semi-automated method utilizing Naïve Bayesian algorithms. The method yielded an overall accuracy of 87% and has significantly reduced the requirement for manual review [7]. The rapid development of ML and natural language processing (NLP) technology has provided a viable path towards automating classification, improving precision, and facilitating continuous surveillance [8].

In another study, an NLP framework was deployed to identify e-scooter-related injuries from over 36 million clinical notes [9]. This NLP approach involved refining regular expression techniques previously developed to identify emergency room (ER) visits related to e-scooter injuries. The framework was trained and tested using a three-stage process on ER visit notes. The training data included notes containing keywords, such as "scooter" and specific e-scooter brand names. The NLP model incorporated an ensemble random multi-model deep learning (RMDL) technique that combined convolutional neural networks (CNNs) and deep neural networks (DNNs) to enhance accuracy. The model used global vectors for word representation (GloVe) and term frequency–inverse document frequency (TF-IDF) to process the text, translating it into a series of vectors that convey semantic meaning and context. The combined scores from the CNN and DNN models were averaged through soft voting to determine the probability of an e-scooter injury. The algorithm indicated an overall accuracy of 92%, correctly classifying the majority of notes in the testing set [9].

In [10], researchers used a comprehensive methodology to predict injury severity using proactive and reactive data from a steel manufacturing plant in India. The dataset was merged to build a mixed dataset, which underwent several preprocessing steps to ensure data quality and manage complexities. The study addressed class imbalance by employing four state-of-the-art oversampling techniques: Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE (BLSMOTE), Majority Weighted Minority Oversampling Technique (MWMOTE), and k-means SMOTE (KMSMOTE). Six classification algorithms were used for injury severity prediction, with KMSMOTE being the most effective in balancing datasets and achieving a higher prediction accuracy. The Random Forest algorithm showed superior performance in terms of accuracy and robustness. The study demonstrated that by integrating proactive and reactive data, and effectively handling class imbalance, advanced ML models significantly improved injury severity prediction, offering valuable implications for safety management and preventive measures.

Van Eetvelde et al. [11] highlighted the growing use of ML in injury prediction and prevention. The researchers reviewed ML techniques, such as tree-based ensemble methods, SVM, and ANN, showcasing that ML models can identify high-risk factors. However, the review called for improved methodological rigor and data analysis and inclusion for optimal performance. These studies did not analyze the performance of the likelihood-based

ensemble approach using combinations of simple and advanced ML models to classify product injury data; hence, in this study, the classification performance of the likelihood-based ensemble ML approach using three well-established ML models, K-Nearest Neighbor, Random Forest, and Logistic Regression, to classify product injury data was evaluated.

3. Methodology

3.1. Data Description

Injury narratives from the National Electronic Surveillance System (NEISS) [2], leveraging their concise and unstructured format, were utilized. The NEISS is an informative dataset managed by the U.S. Consumer Product Safety Commission (CPSC). The system collects data from over 100 hospitals across various healthcare settings, ensuring data representation and an accurate scale to reflect nationwide trends. NEISS is a critical tool for acquiring and assessing information about injuries, including those brought on by defective consumer goods, adverse drug events, and acts of violence. Preventive initiatives, product recalls, and public safety regulations have benefited from the use of NEISS data. The nationally representative hospitals in the United States and its territories feed the data into NEISS. Every ER visit associated with a consumer product or a kid under five years old's poisoning is reported by each member of the NEISS facility [2]. This dataset holds crucial data elements essential for understanding consumer product-related injury dynamics. It offers a demographic profile of those most affected by specific injuries, detailing the age, sex, race, and ethnicity of the injured. The data include detailed injury information, such as descriptions, affected body areas, diagnoses, and outcomes (treated and released or hospitalized), alongside product data highlighting potential risks and identifying involved consumer products. The data used for this research included detailed injury information, such as descriptions, affected body areas, diagnoses, and outcomes, alongside product data highlighting potential risks and identifying involved consumer products.

The NEISS data collection involved the following steps:

(1) User Affiliation: Affiliation was declared by selecting the category "researcher". This classification affects access and offers context for data usage.
(2) NEISS Data Highlights: Access to summary reports of data was granted through this section, with the selection of "Overview-All Products".
(3) Archived Annual NEISS Data: An Excel file was used for storing the raw data acquired for 2022.
(4) NEISS Estimates Query Builder: The NEISS Estimates Query Builder was utilized for specialized data gathering, enabling comprehension of injury codes with few samples (less than 100). Two files, "neiss2022.xlsx" and "DataDictionary042022.xlsx" were retrieved from the NEISS database from their website [2]. The "neiss2022.xlsx" file, which contains three distinct worksheets, namely, "NEISS_2022", "NEISS_FMT", and "Code Details", was utilized. Figure 1 shows a sample record, where the "Narrative_1" column was used as independent variable (X), and the "Diagnosis" column as the dependent variable (Y). The NEISS dataset's injury codes were decoded using the file "DataDictionary042022.xlsx". The document provided the meaning behind each digit in the injury code, including the kind of injury, body part, gender, and product type. Since the dataset was already processed, it did not contain any missing records.

There were a total of 323,344 data points, encompassing 30 unique codes. Table 1 lists the number of injury cases and description for each injury code in the NEISS data used for this research.

Age	Gender	Race	Body Part	Diagnosis	Product	Narrative_1
34	2	1	82	59	478	34YOF WAS WASHING A GLASS THAT BROKE. DX: LACERATION TO RIGHT HAND
11	2	1	82	53	3286	11YOF INVOLVED IN FOUR WHEELER TURNOVER. DX: CONTUSION WITH ABRASION LEFT HAND.
12	2	1	30	53	3286	12YOF WITH MVA, WAS PASSENGER ON A FOUR WHEELER. DX: CONTUSION SHOULDER, CONTUSION RIGHT KNEE.
38	2	1	75	62	698	38YOF, FELL X 4 DAYS GETTING OUT OF HOT TUB, SLIPPED HIT RT SIDE OF HEAD, + LOC DX; FALL, HEADACHE, CLOSED HEAD INJURY, RT KNEE CONTUSION, LT HAND CONTUSION

Figure 1. Sample records in NEISS data files.

Table 1. Data description of different diagnosis codes.

Sr. No	Diagnosis Code	Cause	Data Samples
1	41	Ingestion	3884
2	42	Aspiration	327
3	46	Burns, Electrical	79
4	47	Burns, Not Specified	94
5	48	Burns, Scald	2280
6	49	Burns, Chemical	455
7	50	Amputation	643
8	51	Burns, Thermal	2495
9	52	Concussions	6438
10	53	Contusions, Abrasions	34,775
11	54	Crushing	668
12	55	Dislocation	5113
13	56	Foreign Body	7028
14	57	Fracture	53,849
15	58	Hematoma	4453
16	59	Laceration	50,991
17	60	Dental Injury	1759
18	61	Nerve Damage	1414
19	62	Internal Organ Injury	42,451
20	63	Puncture	2467
21	64	Strain, Sprain	28,840
22	65	Anoxia	940
23	66	Hemorrhage	731
24	67	Electric Shock	192
25	68	Poisoning	4541
26	69	Submersion	380
27	71	Other/Not Stated	61,220

Table 1. *Cont.*

Sr. No	Diagnosis Code	Cause	Data Samples
28	72	Avulsion	2487
29	73	Burns, Radiation	135
30	74	Dermatitis, Conjunctivitis	2214

3.2. Text Preprocessing

The classification of injury narratives requires text preprocessing that converts unstructured text data into a more uniform and analytical structure. Text preprocessing involves steps that are necessary to enhance the interpretability of the text by ML models. The following preprocessing steps were carried out:

3.2.1. Tokenization

Narratives were divided into smaller units, such as words, phrases, and symbols. Analyzing and comprehending the structure and meaning of the text required this distinction. The algorithm took into consideration a number of linguistic quirks and norms, such as how contractions and multi-word sentences were handled.

3.2.2. Cleaning

Characters like punctuation, special symbols, and numerals that do not contribute to the meaning of the text were eliminated. This kind of text cleaning reduced the text's complexity and made it useful for the analysis to concentrate on its important aspects.

3.2.3. Normalization and Stop Words Removal

In order to reduce data dimensionality and the influence of lexical diversity, the text was standardized by changing its case to lowercase. The corpus was restructured by eliminating high-frequency stop words to allow for a more focused analysis on more significant words.

3.3. Training and Test Data

The preprocessed text was split into training and test datasets, following an 80/20 rule to ensure effective training and generalizability. With the allocation of 80% of the data to training, the model learned comprehensively and identified underlying patterns. The remaining 20% was used for testing that assessed the model's ability to generalize to new data. This strategic division facilitated the detection of overfitting and underfitting, enabling tuning to model complexity and training strategy. This approach optimized the ML model's performance across various injury codes and narrative contexts, enhancing its real-world application potential.

3.4. Vectorization

The preprocessed data were transformed into numerical vectors using Count Vectorizer (CV) and Term Frequency Inverse Document Frequency (TF-IDF) Transformer. Count Vectorizer tokenized text narratives by assigning unique features to each word across the corpus. This process enabled ML algorithms to understand the text's content with greater nuance. TF-IDF Transformer refined the textual data representation by calculating token importance based on term frequency within a document and inverse document frequency across the corpus [12]. The integration of CV and TF-IDF transformer accounted for the frequency and uniqueness of words, enhancing document disparity and classification based on thematic relevance [13].

3.5. Machine Learning Models

Three well-established ML models, K-Nearest Neighbor, Random Forest, and Logistic Regression that provided the output in form of prediction and associated likelihood of correctness of prediction, were used.

K-Nearest Neighbor (KNN), a versatile ML model, was utilized for classification tasks. It operates on the principle of instance-based learning, deferring its learning phase until a prediction is required. The core principle of KNN involves predicting the label of a new data point by assessing the "k" nearest labeled data points in the feature space. The "closeness" is determined using a Euclidean distance metric. For this research, KNN was chosen due to its effectiveness in big data analysis, owing to its simplicity and efficiency in handling massive datasets [14]. For each narrative, KNN was utilized to predict both the top 1 and top 2 injury classifications. This approach was adopted to provide a comprehensive understanding of the potential injuries associated with each narrative.

The Random Forest algorithm is an ensemble learning strategy used for classification, which enhances the decision tree method. RF integrates multiple trees to boost prediction accuracy and model stability [15]. The RF model for multiclass classification trains each tree in the forest to predict multiple classes. The training procedure creates bootstrap samples from the original dataset, and each sample is utilized to build a decision tree. RF addresses class imbalance by assigning class weights that effectively prioritize rarer classes. The approach ensures that the less frequented categories are accurately represented in the model's decision-making process.

Logistic Regression effectively extends to multiclass text classification, modeling the text sample's likelihood for predefined categories, and efficiently using the high-dimensional nature of textual data. The "multinomial" method was used to apply LR to injury classification. The approach directly modeled the likelihoods of categories. The model predicted that a given narrative belongs to each code using the SoftMax function, extending the logistic function to multiple categories [16]. This was achieved by computing a set of coefficients for each code that involved maximizing the likelihood of the observed data under the model. These coefficients then determined the influence of each feature on the likelihood of belonging to each category. The sign and magnitude of coefficients indicated the direction and strength of the relationship between features and the log-odds of category elements.

These models were implemented using scikit-learn, a Python library, with default parameters; no hyperparameter tuning was performed.

3.6. Ensemble Modeling

Different ensemble models that integrated the predictive capabilities of the LR, KNN, and RF [17] were developed. The ensemble approach combined the advantageous properties of classifiers into a framework to enhance prediction precision and reliability. The four ensemble models developed were as follows: the KNN + LR model, the KNN + RF model, the RF + LR model, and the KNN + LR + RF model. Combining KNN with LR, we leveraged KNN's ability to capture the nuances of local data structures and LR's powerful probabilistic classification ability. Similarly, making use of RF's strength to handle complex data interconnections and high dimensionality, with KNN's locality-based insights, the number of misclassifications by the individual models were minimized [18].

The third model along with LR used RF, known for its classification efficacy by aggregating multiple decision trees to minimize overfitting while retaining accuracy. The experimental setup and subsequent performance evaluation led to the development of the fourth model [18]. The comprehensive approach integrated the strengths of the KNN, LR, and RF models as depicted in Figure 2. By averaging the likelihood of each category, this model effectively harnessed the combined strengths of the individual models for a more accurate prediction of injury codes from narrative texts. This model enhanced the understanding of injury codes even with limited data points. Unlike conventional strategies that assign a single code to each narrative, these models are intended to predict two po-

tential injury codes, reflecting the complex and diverse character of accidents, which often involve more than one type of injury. The classifier analyzed narratives for probable injury codes that targeted the top 1 (Top1) and top 2 (Top2) potential codes for each narrative. These Top1 and 2 codes were selected based on the likelihood/probability of correctness of prediction, provided as output by the model.

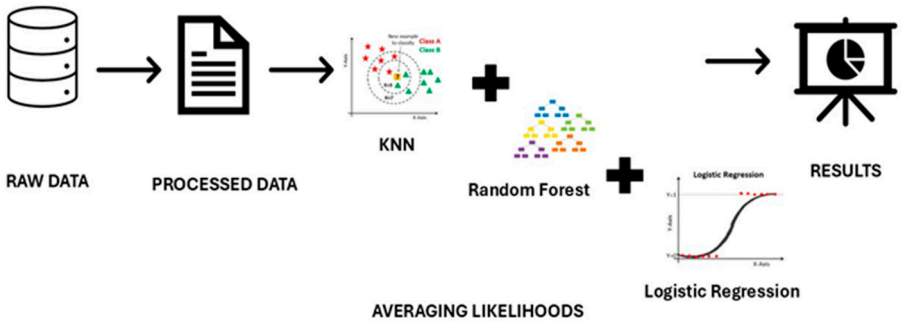

Figure 2. Schematic diagram of ensemble model.

4. Results and Discussion

The performance of these models was quantified using three well-established performance metrics used for machine learning: accuracy, precision, and recall, and the results were reported for both the top 1 (Top1) and top 2 (Top2) predictions. In this context, the Top1 accuracy represents the model's ability to correctly predict the primary injury code, while the Top2 accuracy extends this to include the model's second guess, increasing the chances of a correct prediction. Table 2 presents the performance evaluation metrics for individual ML models of KNN, RF, and LR in predicting the top 1 and 2 injury codes. KNN showed lower performance, particularly in precision and recall for the top 1 predictions, which can be attributed to its sensitivity to noise and reliance on distance measures. In contrast, RF and LR demonstrated superior accuracy in handling high-dimensional data, particularly for the top 2 predictions. LR surpassed RF in precision and recall, benefiting from its strength in binary outcomes and clear decision boundaries.

Table 2. Evaluation metrics of single machine learning models.

Performance Metrics (%)	KNN		RF		LR	
	Top1	Top2	Top1	Top2	Top1	Top2
Accuracy	71.3	83.6	90.2	97.5	92.2	98.6
Precision	78.2	51.5	89.0	55.5	91.6	55.2
Recall	65.4	77.1	64.9	81.2	77.9	89.7

The RF model showed superior performance in the Top1 predictions across all metrics, particularly in accuracy (90.19%) and precision (89.02%), indicating its effectiveness in correctly identifying the primary injury code with high reliability. The LR model, however, showed remarkable results in the Top2 accuracy (98.63%), suggesting its strength in providing a broader set of potential injury codes that likely include the correct one. Table 2 underscores the variability in model performance based on the metric considered, highlighting the need to balance precision (the model's ability to return relevant results) and recall (the model's ability to find all relevant instances) based on the application's requirements.

The detailed performance of the LR model for each category is shown as a confusion matrix in Figure 3. It depicts the LR model's performance in multi-class variation, highlighting the model's challenges in correctly classifying complex data due to high prediction errors and misclassifications. The results indicate that using a single model, such as LR, is ineffective for handling complex datasets, thus necessitating the need for ensemble models,

which combine the strengths and capabilities of multiple algorithms to improve prediction accuracy and reduce errors.

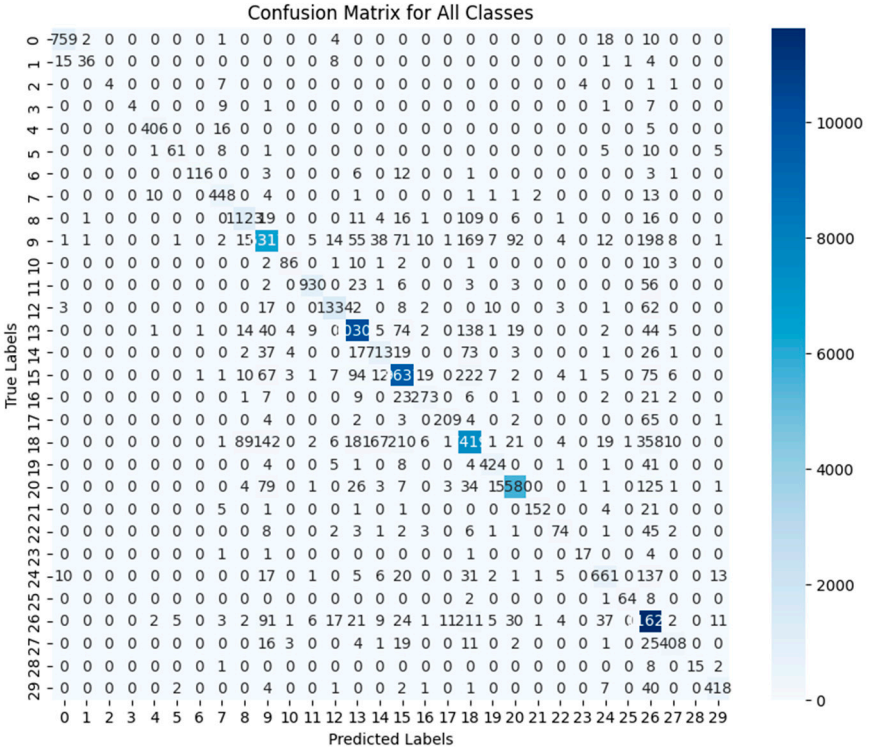

Figure 3. Confusion matrix of top-1 predictions of LR model.

Tables 3 and 4 indicate the analysis of ensemble models, which combine predictions from multiple ML techniques to improve overall prediction accuracy. Table 3 indicates the Top1 prediction performance of ensemble models made from different combinations of KNN, RF, and LR. The RF + LR ensemble showed the highest accuracy (92.27%) and precision (92.01%), suggesting that this combination is particularly effective at identifying the primary injury code accurately. The combination of all three models (KNN + RF + LR) proved to be the most accurate, demonstrating a significant improvement in predictive modeling. The differences in precision and recall among various combinations (e.g., KNN + LR, RF + LR) highlighted the unique advantages of each ensemble, owing to their strategic integration of base model capabilities. Compared to Table 2, showing that individual models had distinct performance limitations, the shift to ensemble methods was an effective strategy to overcome these challenges. While Table 2 presents the highest accuracy of 92.19% (LR for top 1) and 98.63% (LR for top 2), the peak ensemble accuracy in Table 3 is 97.36%. This improvement of up to 5.17% in the top 1 prediction confirms the effectiveness of combining different modeling techniques to address complex prediction tasks.

Table 3. Evaluation metrics of ensemble models—Top one prediction

Performance Metrics (%)	Ensemble Models			
	KNN + LR	KNN + RF	RF + LR	KNN + RF + LR
Accuracy	90.47	83.88	92.27	97.36
Precision	89.66	86.54	92.01	89.08
Recall	77.50	70.14	75.99	76.99

Table 4. Evaluation metrics of ensemble models—Top two predictions.

Performance Metrics (%)	Ensemble Models			
	KNN + LR	KNN + RF	RF + LR	KNN + RF + LR
Accuracy	97.12	95.16	98.50	99.50
Precision	53.04	53.30	59.50	60.23
Recall	88.54	83.58	88.04	88.03

Table 4 depicts the effectiveness of ensemble models in predicting the top two injury codes, representing a trend of improved accuracy and precision compared to single-model predictions. It shows that the ensemble methods address the limitations observed in single models within complex scenarios, as highlighted in Table 2. The KNN + RF + LR ensemble model excelled in accuracy, demonstrating its ability to integrate a broader range of relevant features for more precise predictions. The ensemble model had a drop in precision, from a peak of 91.51% (LR in Table 2 for top 1) to 60.23%. Despite this, the high recall rate compensated for the precision decline, indicating that while the ensemble predicted more than one code, it reliably identified the correct ones due to its comprehensive approach. By integrating the diverse strengths of KNN, RF, and LR, the ensemble model effectively navigated the complexities of the data, ensuring that relevant predictions were not missed.

Figure 4 shows the confusion matrix, illustrating the performance of the ensemble model with KNN, RF, and LR. The values along the diagonal are closer to zero, indicating more accurate predictions compared to the LR model alone. Misclassifications, represented by off-diagonal values, are reduced, suggesting that the ensemble model handles the data efficiently, overall, improving the accuracy and reducing extreme prediction errors.

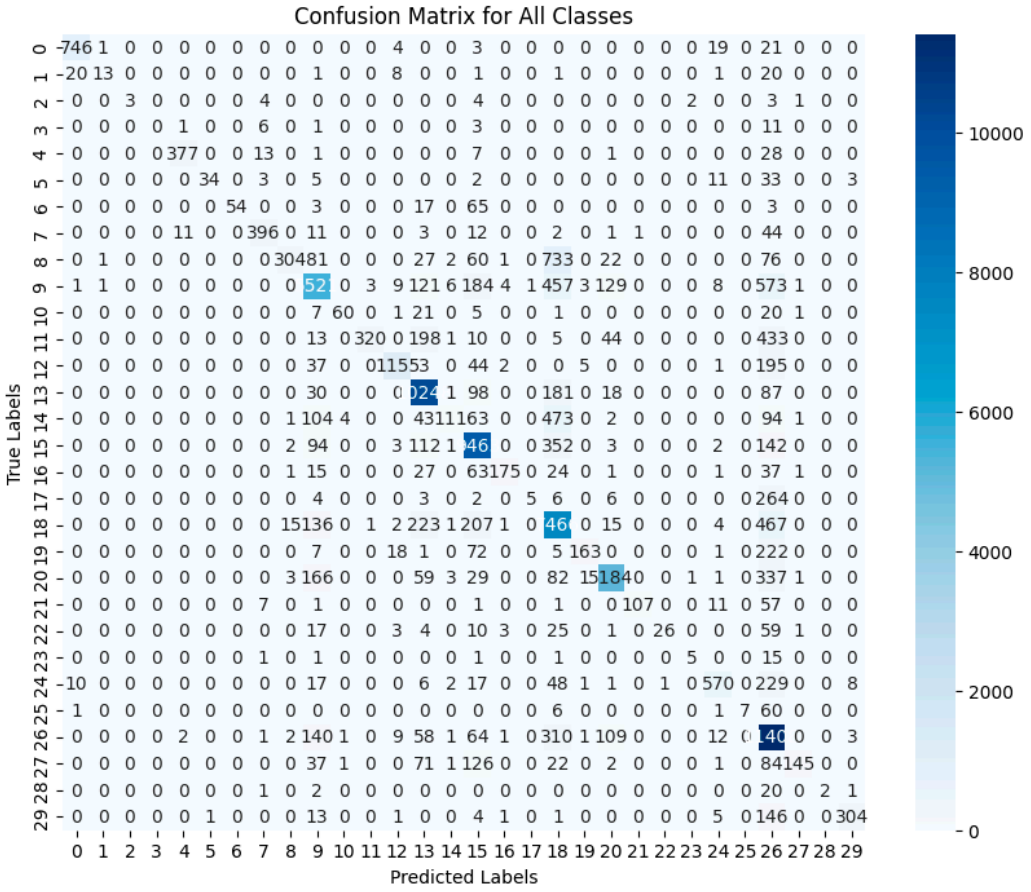

Figure 4. Confusion matrix for top-1 predictions of ensemble model KNN + RF + LR.

5. Conclusions

We developed ensemble models that enhanced the predictive performance across a number of criteria by combining three of the most popular ML models: K-Nearest Neighbor, Random Forest, and Logistic Regression. Our results indicate that ensemble models, especially the KNN +RF + LR combination, are highly effective for injury code prediction using NEISS data, providing a robust tool for accurately identifying injury types from narrative reports. The inclusion of both the Top1 and Top2 prediction evaluations offers a comprehensive view of each model's performance, emphasizing the trade-offs between capturing the most likely injury code and ensuring broader coverage to include possible alternatives.

The combined KNN, LR, and RF model indicated an accuracy of as high as 99.50% for the top two injury predictions, while the KNN + LR ensemble model showcased a robust predictive power with a high accuracy of 90.47%. Our results demonstrate that ensemble machine learning is a useful tool for enhancing clinical accident narrative classification, particularly when individual ML models do not yield a very high accuracy. Future research in clinical narrative analysis can utilize the presented ensemble ML models framework for decision support in injury surveillance and prevention.

The proposed ensemble models' ability to make multiple predictions enhanced reliability and also ensured that cases with fewer samples were considered, which might be overlooked by a conventional (single) prediction model due to underfitting. In real-world applications, this is particularly valuable because certain injury types or scenarios may be underrepresented in the data, making it challenging for a single model to make accurate predictions. The ensemble models address this issue by combining the strengths of multiple algorithms, ensuring that even rare or less common cases are captured in predictions. The ensemble model approach can, thus, aid in emergency response by accurately predicting and prioritizing critical injuries that are not frequently reported, ensuring appropriate resources are allocated in real-world applications.

Our findings demonstrate the potential of ensemble models to mitigate data scarcity, laying the groundwork for methodological advancements. These probability-based ensemble models can be deployed to predict injury trends, identify emerging risks, and improve workplace safety. This application of AI-driven text analysis enables a comprehensive understanding of injury causation and will help in informed preventive measures, thus promoting safety and public health on a broader scale. This approach can also be applied to other predictive modeling scenarios, where data imbalance is an issue, ensuring that cases with lower sample sizes and less frequent occurrences are given appropriate consideration. The proposed ensemble model has potential for advancing predictive analytics in healthcare, public health, and high-risk industries.

Author Contributions: Conceptualization, S.V., G.N. and R.S.; methodology G.N. and S.V.; software, S.V.; validation, S.V., G.N. and R.S.; resources, G.N. and R.S.; data curation, S.V.; writing—original draft preparation, S.V.; writing—review and editing, G.N. and R.S.; visualization, S.V.; supervision, G.N. and R.S.; project administration, G.N. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data used is publicly available on the NEISS website [2].

Conflicts of Interest: The authors declare no conflict of interest

References

1. Cohan, A.; Fong, A.; Ratwani, R.M.; Goharian, N. Identifying harm events in clinical care through medical narratives. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20–23 August 2017. [CrossRef]
2. CPSC NEISS On-Line Query System. U.S. Consumer Product Safety Commission. Available online: https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx (accessed on 10 October 2023).
3. Williamson, A.; Feyer, A.-M.; Stout, N.; Driscoll, T.; Usher, H. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. *Inj. Prev.* **2001**, *7*, i15–i20. [CrossRef] [PubMed]
4. Sandelowski, M. Telling Stories: Narrative approaches in qualitative research. *Image-J. Nurs. Scholarsh.* **1991**, *23*, 161–166. [CrossRef] [PubMed]
5. Marucci-Wellman, H.R.; Lehto, M.R.; Corns, H.L. A combined Fuzzy and Naive Bayesian strategy can be used to assign event codes to injury narratives. *Inj. Prev.* **2011**, *17*, 407–414. [CrossRef] [PubMed]
6. Wellman, H.; Lehto, M.R.; Sorock, G.S.; Smith, G.S. Computerized coding of injury narrative data from the National Health Interview Survey. *Accid. Anal. Prev.* **2004**, *36*, 165–171. [CrossRef] [PubMed]
7. Marucci-Wellman, H.R.; Corns, H.L.; Lehto, M.R. Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. *Accid. Anal. Prev.* **2017**, *98*, 359–371. [CrossRef] [PubMed]
8. Gasparetto, A.; Marcuzzo, M.; Zangari, A.; Albarelli, A. A survey on Text Classification Algorithms: From Text to Predictions. *Information* **2022**, *13*, 83. [CrossRef]
9. Ioannides, K.L.; Wang, P.-C.; Kowsari, K.; Vu, V.; Kojima, N.; Clayton, D.; Liu, C.; Trivedi, T.K.; Schriger, D.L.; Elmore, J.G. E-scooter related injuries: Using natural language processing to rapidly search 36 million medical notes. *PLoS ONE* **2022**, *17*, e0266097. [CrossRef] [PubMed]
10. Sarkar, S.; Pramanik, A.; Maiti, J.; Reniers, G. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and Reactive Data. *Saf. Sci.* **2020**, *125*, 104616. [CrossRef]
11. Van Eetvelde, H.; Mendonça, L.D.; Ley, C.; Seil, R.; Tischer, T. Machine learning methods in sport injury prediction and prevention: A systematic review. *J. Exp. Orthop.* **2021**, *8*, 27. [CrossRef] [PubMed]
12. Zhang, Y.; Gong, L.; Wang, Y. An improved TF-IDF approach for text classification. *J. Zhejiang Univ.* **2005**, *6*, 49–55. [CrossRef]
13. Gupta, A.; Sharma, U. Machine Learning based Sentiment Analysis of Hindi Data with TF-IDF and Count Vectorization. In Proceedings of the 2022 7th International Conference on Computing, Communication and Security (ICCCS), Seoul, Republic of Korea, 3–5 November 2022; pp. 1–5. [CrossRef]
14. Deng, Z.; Zhu, X.; Cheng, D.; Zong, M.; Zhang, S. Efficient kNN classification algorithm for big data. *Neurocomputing* **2016**, *195*, 143–148. [CrossRef]
15. Wu, Q.; Ye, Y.; Zhang, H.; Ng, M.K.; Ho, S.-S. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowl.-Based Syst.* **2014**, *67*, 105–116. [CrossRef]
16. Ramadhan, W.P.; Astri Novianty, S.T.M.T.; Casi Setianingsih, S.T.M.T. Sentiment analysis using multinomial logistic regression. In Proceedings of the 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), Yogyakarta, Indonesia, 26–28 September 2017; pp. 46–49. [CrossRef]
17. Shah, K.; Patel, H.; Sanghvi, D.J.; Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.* **2020**, *5*, 12. [CrossRef]
18. Vichare, S.S. Probabilistic Ensemble Machine Learning Approaches for Unstructured Textual Data Classification. Master's Thesis, Purdue University Graduate School, West Lafayette, IN, USA, 26 April 2024.