# Comparing Violence Detection in Videos Using Classification Algorithms

Alicia Unterreiner

Department of Computer Science, Binghamton University

CS301: Ethical, Social, and Global Issues in Computing

Dr. George Weinschenk

May. 10, 2023

**Abstract**

Leading to a decline in mental health, sexual violence on college campuses poses a threat to students. However, the detection of the physical violence that perpetrators of sexual assault utilize to subdue victims may help alert law enforcement to the crime. This paper presents a comparison of the classification algorithms: Convolutional Neural Network (CNN) and Dempster-Shafer (D–S). The CNN algorithm analyzes a video displaying violence using auditory or visual components while the D–S uses both visual and auditory components to determine the presence of violence. Convolutional neural networks process images by applying filters to an input to produce a feature map: a summary of the identification of features such as violence. The D–S algorithm employs the Dempster–Shafer theory of uncertainty with convolutional neural networks to combine the detection of audio and images. Since the CNN algorithm compares visual and audio components in video separately, evidence conflict bars a conclusive result. The D–S algorithm improves upon the existing Dempster–Shafer theory by adjusting basic probability assignment functions to avoid veto power and by utilizing conflict levels to determine the validity of conclusions. The improvements significantly increase the accuracy of the algorithm. The faster the detection of assault, the faster law enforcement can stop the crime which highlights the importance of accuracy in the detection of sexual violence. Therefore, the Dempster–Shafer (D–S) classification algorithm performs better in detecting violence through video. Future work includes violence detection utilizing videos of sexual assault, instead of videos of general violence discussed in this paper.

## Comparing Violence Detection with Audio Using Classification Algorithms

Sexual violence on college campuses can lead to depression and suicidal thoughts in victims. Physical violence to subdue victims of sexual assault by the perpetrator and the assault at times remains undetected until the attack concludes. To better detect sexual violence, researchers use classification algorithms on campus surveillance tapes to flag images and sounds found in situations involving sexual violence. The Dempster–Shafer (D–S) classification algorithm in comparison to the Convolutional Neural Network (CNN) classification algorithm performs better in detecting violence through video to allow law enforcement to find and stop attacks quicker. CNN classification algorithms use matrix filtering to detect shapes in images. When used in audio, the CNN algorithm identifies emotions associated with violence by performing classification on mel-frequency cepstral coefficients (MFCCs). MFCCs describe a representation of the sound power spectrum. To solve the problem of conflicting evidence when using the classification results of physical violence and speech recognition, researchers present the Dempster-Shafer classification algorithm. The combination of audio and image detection in the improved Dempster-Shafer classification algorithm yields high accuracy in identifying violence. The basis of the D-S algorithm relies on the CNN algorithm which previously existed as the most accurate classification algorithm.

## Alternative Technology

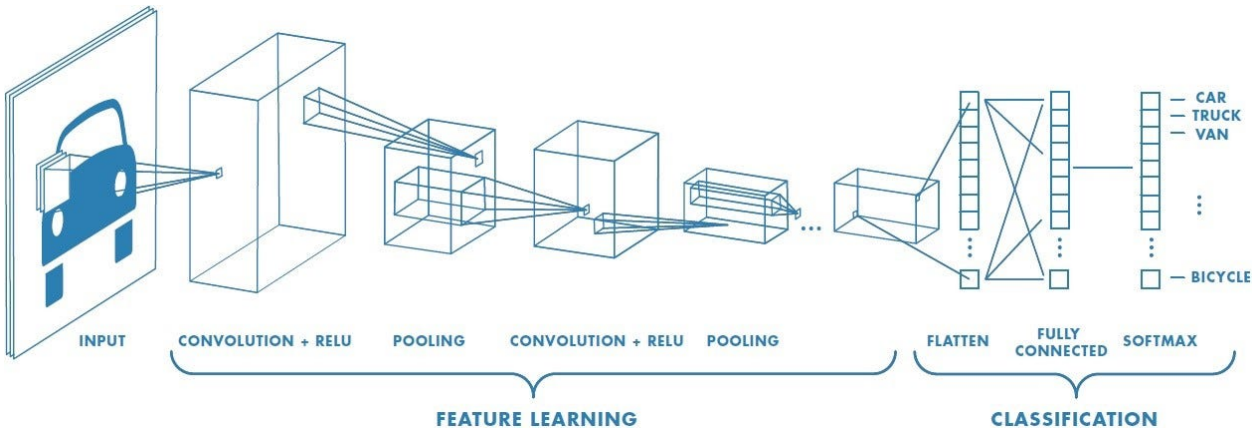
Neural networks, a type of machine learning called deep learning, facilitate the analysis of information. According to IBM Cloud Education (2020), node layers containing additional layers compromise neural networks (p. 2). These nodes work by passing data to continual layers in the network given the output of a node exceeds a maximum value. Previous to the existence of CNNs, the identification of images remained a manual and therefore a slow task (IBM Cloud Education, 2020, p. 1). Research on CNNs by Kunihiko Fukushima and Yann LeCun created a base of knowledge on the topic in 1980 (IBM Cloud Education, p. 3). LeCun utilized neural networks to find patterns within zip codes. Presently, convolutional neural networks provide large-scale pattern recognition.

The Convolutional Neural Network (CNN) — a type of artificial neural network — differs from other neural networks in the network's three main types of layers: convolutional layer, pooling layer, and

fully connected (FC) layer (p. 2). The article notes that the convolutional layer stands as the main component of a CNN and exists as the location of matrix multiplication. The components of the convolutional layer include input data, a filter, and a feature map (p. 2). Figure 1 shows the process called a convolution used by the algorithm which works by a filter moving across input data — such as an image — to verify the presence of a pattern or shape (p. 3). IBM explains that the pooling layer reduces complexity and time-consumption by minimizing the number of input parameters while the fully-connected layer contains nodes that connect to previous nodes in other layers. The complexity of the CNN grows as the number of layers increases with later layers focusing on complex features until the algorithm finds the object (p. 2).

**Figure 1**

*Diagram Displaying the Recognition Process of a Convolutional Neural Network*




The structure of CNNs allows the algorithm to process complex images efficiently and accurately. The study by Liang Ye et al. (2021) — researchers from multiple universities with varying levels of academic degrees in computer science —found that the use of a 3D CNN which utilizes a three-dimensional filter for audio classification achieved a recognition accuracy of 92.00%; thus, the emotions identified by the algorithm matched the emotions portrayed in the video (p.1). To classify the emotions, the research utilized three speech emotion databases. The Dempster–Shafer (D–S) classification

algorithm exists as the better choice in the algorithm for the identification of campus violence due to the limited accuracy of the audio recognition in the CNN classification algorithm.

## Support

### Technical Details

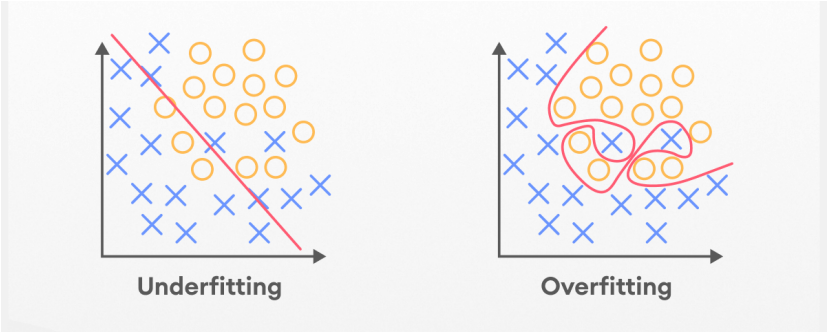
The Dempster-Shafer (D–S) classification algorithm extracts features from video and audio to complete an assessment of video recordings. To analyze images from videos, a C3D neural network uses maximum pooling and convolution operations. Brownlee (2019) — a machine learning specialist with a Ph.D. in machine learning from Swinburne University of Technology — describes maximum pooling operations as calculations to find the largest value in each feature map patch, a part of the output achieved from applying filters to the input (p. 3). To produce a feature map convolution operations filter the information in 3-directions for the C3D neural network (Ye et al., 2021, p.3). After the operations, the neural network includes two connected layers with 4096 neurons in the first layer and 487 in the second layer (p. 3). An additional 4-layer neural network receives input from the 4096-neuron feature vector in the C3D neural network (p. 3). Tyka (2016), a computational biophysics and biochemistry researcher at the University of Washington, describes neurons in the feature vector as units that compute simple calculations and transfer the results of the calculations to neurons in the network (4:41).

Although neural networks learn from data by finding patterns in images overfitting presents an issue. IBM Cloud — a computing service offered by the company IBM — introduces that overfitting occurs when a statistical model matches the data used to train the model (IBM, 2020, p. 1). As seen in Figure 2 when overfitting occurs, the algorithm learns irrelevant information within the training data and only can generalize to that data, thus defeating the algorithm's ability to predict or generalize new information (p. 1). Sagar (2019), a machine learning researcher at the University of Maryland, writes that ways to avoid overfitting include: simplifying the training model, preventing the neural network from learning from parts of the model, augmenting the data, and using dropouts. The algorithm uses dropout in hidden layers in the 4-layer neural network to prevent overfitting. Dropouts work to make the model less

complex by modifying the network (para. 5). To modify the network, neurons get randomly removed during the training process which equates to training a variety of networks (para. 5).

**Figure 2**

*Underfitting Lowers the Ability of a Model to Make Conclusions and Generalizations*




To train the model, the Adagrad optimizer — which has adaptive learning rates — undergoes an iterative process (Brownlee, 2021, para. 7-8). The optimizer randomly selects sample data and the samples' labels and calculates the value and error of the gradient, a derivative vector that indicates differences. The optimizer changes the variable corresponding to the sum of the gradient values and modifies the parameters relative to the values (para. 9). The article explains that after the compilation of predicted results based on the inputs, a conversion of the results into probability vectors occurs for further analysis. The usage of mel-frequency cepstral coefficients (MFCC) helps to detect emotions in audio.

According to Miroslaw Plaza et al. (2022), computer science researchers from multiple universities, identifying emotions using a computer presents a difficult task due to factors such as the availability of algorithms to recognize emotions (para. 5). Veloso et al. (2022), researchers in the field of computer science from the University of Minho and the University of Porto, state the benefit of the representing of emotions in audio when utilizing convolutional neural networks (para. 1). Inputting MFCC into a deep learning framework called Kera to detect emotions in audio. Craig Smith (2022) — the founder of a podcast focused on artificial intelligence — introduces Keras as a framework that dominates the artificial intelligence field. Obtaining MFCC from audio using Keras involves isolating the parts of the audio that contain human speech using voice activity detection (Ye et al., 2021, p. 4) A fast Fourier

transform then converts the time domain, the variation of amplitude, of the separated signals into a function of frequency (p. 4). To produce the MFCCs, the frequency domain signals went through Mel filters and discrete cosine transform (DCT) (p. 4). On the point of discrete cosine transforms, Tanveer Singh (2019), a software engineer in the signal processing field, explains DCT as a way to decorrelate overlapping signals and find the cepstral coefficients (p. 2).

The Keras framework uses the MFCC features to output a probability of true or false depending on the detection results. A total of 6 convolution layers and 2 max-pooling layers make up the neural network (Ye et al., 2021, p. 4). Gradient vanishing, explained by Wang (2019) — a computer science major with past internships at Facebook and Meta — describes when the gradients of the loss function approach zero which makes the gradient too small for effective training pose a problem for data analysis (para. 1). The issue of gradient vanishing becomes a severe problem the more layers exist in the network (para. 4). If given negative input the rectified linear unit (ReLU) returns 0 to avoid gradient vanishing (Ye et al., 2021, p. 4). Wang explains that similar to the neural network used for visual recognition, a dropout layer prevents overfitting and a flattened output layer gives the results of the analysis.

To combine the results of the video and audio emotion procedures, the Dempster-Shafer reasoning theory considers the outcomes of a problem depending on the evidence given (p. 5). However, D–S harbors the weaknesses: if a heavy conflict exists, the result lacks accuracy and the probability distribution function biases the result (p. 5). Adjustment of the basic probability assignment and taking conflict levels into account improves the D–S fusion theory by increasing the accuracy of the results (p. 5).

The new D–S classification algorithm redefines basic probability assignment (BPA) functions to avoid the problem of veto power (p. 6) . The first step of the D–S algorithm, basic probability assignment, according to Jiang et al (2016), researchers in engineering, lacks a requirement of as many conditions as probability theory and includes the variable of uncertainty (para. 1). By using evidence from test scenarios in calculations, BPA represents both the results of the research and how uncertainty factors into the results. An inherent problem in BPA functions, veto power, describes when a BPA of evidence calculates

to zero, a negation of the hypothesis occurs (p. 6). To use combinations of video and audio samples, veto power must not exist in the computation. To solve the problem, the improved Dempster-Shafer (D–S) classification algorithm modifies BPA functions to exist in exponential form instead so that the BPA function fails to calculate to 0 (Ye et al., 2021, p. 6). In addition to the redefinition of BPA functions, the Yager fusion rule improves the D–S algorithm by using evidence credibility to avoid evidence conflict (p. 7). Evidence credibility calculates the level of support and credibility of the evidence.

With the improvements made to the Dempster-Shafer algorithm that combined the results of enhanced video and audio detection, the accuracy of the algorithm reached 97.00% with a precision of 97.96% (p. 8). The accuracy and precision of the new algorithm supersede the existing D–S theory with an increase in recognition accuracy by 10.79% (p. 8). With such high accuracy, the new D–S algorithm stands to make a large impact when applied to technologies used in daily life.

**Social Impact**

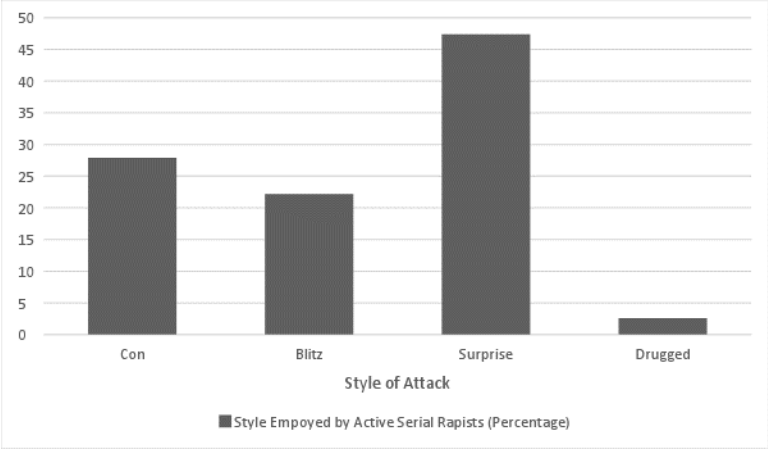
The utilization of an improved Dempster-Shafer algorithm stands to increase the accuracy of violence detection, significantly impacting the safety of citizens. ADT — a company that provides security services — declared that after the installation of surveillance cameras in Orange County, New Jersey, crime dropped by 50% (para. 6). The decrease cited by ADT highlights the importance of surveillance tapes in keeping crime down and citizens safer. Additionally, ADT notes that police officers said the surveillance system aided in investigating criminal cases (para. 6). The ability to see the crimes committed on video with audio and visual representation of the crime contributes significant detail in reporting the violence or criminal nature of the act. The recognition of crimes on surveillance videos allows officers to solve cases after the fact according to ADT and at the time of the crime. With the D–S algorithm to recognize violence, alerts sent out to the police based on the results of the algorithm stop violence and reduce harm to citizens.

Halting violence early ends crimes before the acts cause more harm to the victim. In the case of sexual assault crimes, according to Hazelwood (1990), an FBI agent who profiled sexual crimes, in the blitz approach, the assailant physically assaults the victim (para. 1). Figure 3 includes this style of attack

as one of the styles most used by rapists which highlights the prevalence of the blitz attack. The assault subdues the victim which allows the perpetrator to carry out a crime of sexual violence without resistance or harm to the perpetrator by the victim (para. 1). Surveillance cameras that automatically detect violence using the D–S algorithm stop perpetrators of violence in general and perpetrators of rape who use the blitz approach through alerting police who act to protect the victims. With a lack of violence detection during the committing of a crime, a rapist who uses the blitz approach goes undetected. After the overt violent portion of the blitz attack, the assailant remains free to continue to attack the victim sexually. The acknowledgment that sexual violence lacks the overtness a violent attack possesses due to sounds made in resistance by the victim that cease after the subduing of the victim. Detecting violence stops sexual assault sooner and saves the victim from trauma caused by the sexual assault and the consequences of that trauma.

**Figure 3**

*Styles of Attack Used by Rapists Include Con, Blitz, Surprise, and Drugging*

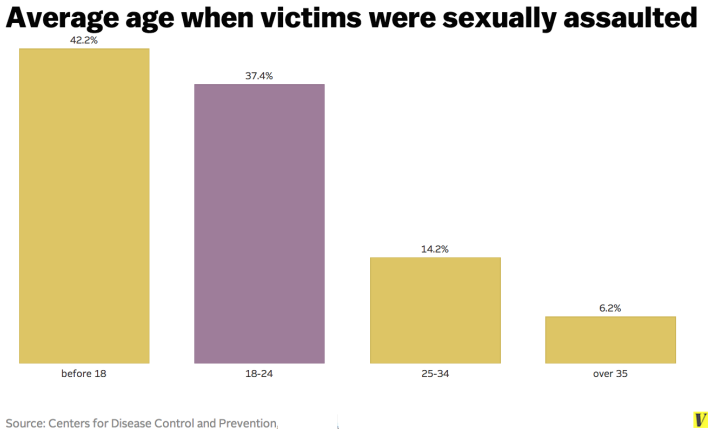



The prevalence of sexual assault in America, especially on college campuses, proves the importance of algorithms with the accuracy to detect the precursors of sexual violence. Figure 4 displays the second most amount sexual assault victims at ages 18-24, a time when victims may enroll in college. The National Sexual Violence Resource Center (2015), an organization that promotes the prevention of sexual violence, writes that in college, sexual assault happens to 1 in 5 women and 1 in 16 men (para. 4).

With sexual violence, consequences to the victim include depression, post-traumatic stress disorder, and suicidal ideations. When sexual assault lacks prevention strategies, there exists an increase in the number of victims suffering as a result of the attack. The responsibility of protection lies in a college's duty to those enrolled. The high rate of sexual assault on campuses brings to attention that colleges must take action and use detection strategies with high accuracy to detect violence such as the D–S algorithm. The greater the accuracy of violence detection, the greater the detection and halt of attacks.

**Figure 4**

*Likelihood of Sexual Assault Second-Highest When the Age of Victims Ranges From 18-24*




## Conclusion

The Dempster-Shafer (D–S) algorithm provides more benefits to society than the Convolutional Neural Network (CNN) algorithm. When provided with both auditory and visual inputs, the CNN algorithm performs worse than the D–S algorithm due to evidence conflict. The CNN algorithm processes audio and images separately whereas the D–S algorithm processes audio and images together with the consideration of evidence conflict. With this consideration, the D–S algorithm evaluates the presence of violence in video using audio and images with a low amount of uncertainty. Higher accuracy in the detection of violence means the ability to alert law enforcement to cases of sexual assault that may normally go unnoticed. The utilization of the Dempster–Shafer (D–S) classification algorithm results in higher accuracy of violence detection than the Convolutional Neural Network (CNN) classification

algorithm, giving law enforcement the ability to halt sexually violent crimes and therefore prevent mental health decline of possible victims. Academic professionals on college campuses should recognize the importance of detecting sexual crimes committed against students. The college's responsibility includes keeping the students on the college campus safe and the algorithm with the greatest accuracy can lead to an increase in the safety of students. Therefore, computer professionals should choose the Dempster-Shafer (D–S) algorithm for security systems put in place to detect violence to protect students. The findings regarding the D–S algorithm indicate the importance of future research to improve the capacity of the algorithm to detect sexual assault. Since the findings of the paper revolve around the recognition of general violence, the next step involves detecting specific cases of sexual violence in video. Future studies should use videos that cover stages of a sexual crime in the determination of the presence of sexual assault. Stages include an assailant making unwanted advances on the victim, physical violence that precedes sexual violence, and sexual assault. The analysis of each stage separately and then the next stage combined with the previous, establishes a warning system for when sexual assault may occur and an alert system for an active sexual assault. With a lack of a system in place to detect the crime of sexual violence, sexual assault continues rampant on college campuses, harming the mental health of students who will become the future of America.

## References

Brownlee, J. (2019, July 5). *A gentle introduction to pooling layers for convolutional neural networks*. Machine Learning Mastery. https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/

Brownlee, J. (2021, October 11). *Gradient descent with Adagrad from scratch*. Machine Learning Mastery. https://machinelearningmastery.com/gradient-descent-with-adagrad-from-scratch/

*Do surveillance cameras deter crime?* ADT Security Systems. (n.d.). https://www.adt.com/resources/do-surveillance-cameras-statistically-reduce-crime

Fesmire, C., Vander Ven, T., & Wright, L. (2019). The social camouflage and everyday masks of the con-style serial rapist: a sociological analysis of newspaper accounts. [Graph]. QualitativeCriminology. https://www.qualitativecriminology.com/pub/v7i3p2/release/1

Hazelwood, R., & Warren, J. (1990). Criminal behavior of the serial rapist. *Polygraph*, *19*, (2), 139-146. https://www.ojp.gov/ncjrs/virtual-library/abstracts/criminal-behavior-serial-rapist-0#:~:text=In%20the%20blitz%20approach%2C%20the,her%20after%20she%20is%20sleeping

IBM Cloud Education. (2020, October 20). *What are convolutional neural networks?* https://www.ibm.com/cloud/learn/convolutional-neural-networks

Jiang, W., Zhan. J., Zhou, D., & Li, X. (2016, May 24). A method to determine generalized basic probability assignment in the open world. *Mathematical Problems in Engineering, 2016.* 10.1155/2016/3878634

*Media packet: statistics about sexual violence*. National Sexual Violence Resource Center. (2015). https://www.nsvrc.org/publications/nsvrc-publications-fact-sheets/media-packet-statistics-about-sexual-violence

Nelson, L. (2014). Six charts that explain sexual assault on college campuses. [Graph]. Vox. https://www.vox.com/2014/5/9/5696162/6-facts-about-sexual-assault-on-campus

Płaza, M., Trusz, S., Kęczkowska, J., Boksa, E., Sadowski, S., & Koruba, Z. (2022). Machine learning algorithms for detection and classifications of emotions in contact center applications. *Sensors*, *22*, (14), 5311. 10.3390/s22145311

Sagar, A. (2019, December 6). *5 techniques to prevent overfitting in neural networks*. KDnuggets. https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html

Saha, S. (2018). A comprehensive guide to convolutional neural networks — the ELI5 way. [Image]. SaturnCloud. https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/

Singh, T. (2019, June 18). *MFCC's made easy*. Medium. https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040

Smith, C. S. (2023, March 29). *Baidu's PaddlePaddle spins AI up to industrial scales*. IEEE Spectrum. https://spectrum.ieee.org/paddlepaddle-baidu

Tigran, P. (2022). Overfitting and underfitting in machine learning. [Image]. SuperAnnotate. https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/

Tyka, M. (2018). *The art of neural networks* [Video]. TEDxTUM. Mike Tyka: The art of neural networks | TEDxTUM

Veloso, B., Durães, D., Novais, P. (2022, October 13). *Analysis of machine learning algorithms for violence detection in audio*. Highlights in Practical Applications of Agents, Multi-Agent Systems, and Complex Systems Simulation. The PAAMS Collection. PAAMS 2022. Paper presented at International Conference on Practical Applications of Agents and Multi-Agent Systems, Location (210-221). *Communications in Computer and Information Science*, *vol 1678*. Springer, Cham. 10.1007/978-3-031-18697-4_17

Wang, C.-F. (2019, January 8). *The vanishing gradient problem*. Medium. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

*What is overfitting?* IBM. (n.d.). https://www.ibm.com/topics/overfitting

Ye, L., Liu, T., Han, Tian., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2021, February 9). Campus violence detection based on artificial intelligent interpretation of surveillance video sequences. *AI Interpretation of Satellite, Aerial, Ground, and Underwater Image and Video Sequences*, *13*. 10.3390/rs13040628