



Developing a standardized approach for bite mark identification: introducing a Principal Component Analysis based framework for dental impressions

Yilmaz B

Submitted: November 4, 2024, Revised: version 1, January 13, 2025

Accepted: January 20, 2025

Abstract

Forensic odontology plays a crucial role in criminal investigations and victim identification, especially when other evidence is unavailable. Bite mark analysis, a key area within this field, has long been criticized for its lack of standardization and susceptibility to examiner bias. This study addresses these challenges by introducing a standardized methodology that uses Principal Component Analysis (PCA) to compare bite marks using 3D dental molds. By developing a composite scoring system that captures key dental features, such as inter-molar distances, surface areas, and rotational angles, this approach reduces subjective variability, aiming to establish a scientifically rigorous framework for bite mark comparisons. Dental impressions from 30 participants (all aged 16 to maintain age consistency) were analyzed through high-resolution imaging and statistical methods. Key performance metrics include a precision of 0.67, recall of 0.69, specificity of 0.95, accuracy of 0.91, an Area Under the Curve (AUC) of 0.91, and a Precision-Recall Area Under the Curve (PR-AUC) of 0.75. These metrics highlight the model's ability to balance precision and recall while minimizing both false positives and false negatives—essential requirements in forensic contexts where reliable evidence is critical for accurate legal outcomes. Compared to existing studies in forensic odontology, this model's precision and AUC are consistent with or exceed those reported in similar quantitative methodologies for bite mark analysis. However, similar to prior research, this study highlights the importance of larger and more diverse datasets to validate and extend its applicability. Expanding demographic diversity and testing across different forensic scenarios will be essential for broader implementation. This research presents an advancement in the standardization of forensic bite mark analysis, laying the groundwork for future studies to refine the model and enhance its reliability. By establishing a data-driven approach, this study contributes toward improving the credibility and reproducibility of forensic evidence.

Keywords

Forensic odontology, Bite mark analysis, Principal Component Analysis, Dental impression, Composite scoring, Forensic science, Evidence standardization, Cosine similarity, Vector similarity, Imbalanced data

Berke Yilmaz, Tesla STEM High School, 4301 228th Ave NE, Redmond, WA 98053, USA.

berkeyilmaz.us@gmail.com

1 Introduction

1.1 Background

Forensic odontology, a specialized branch of forensic science, plays a crucial role in the identification of individuals by an examination of their dental records, particularly in cases where traditional methods; such as fingerprints; are unavailable. Forensic odontologists use various forms of dental evidence, including bite marks, restorations, and anatomical characteristics, to link individuals to crime scenes or assist in the identification of victims in mass disasters (1). The field has grown in significance because it enables identification even when human remains are incomplete, damaged, or degraded (2).

Relying on the unique features of an individual's dentition; including tooth number, position, condition, and dental work such as fillings or prosthetics; forensic odontology proves invaluable in mass disaster scenarios where traditional soft tissue identification is not feasible (3). For instance, dental records have proven reliable in many investigations involving severely disfigured remains (4).

The pioneering work of Dr. Oscar Amoedo, widely recognized as the father of forensic odontology, laid the foundation for the use of dental records in legal investigations. His work in identifying victims of a fire accident in Paris in 1898 has continued to influence modern forensic practice (3). Since then, the field has become an invaluable tool in identifying victims of disasters like earthquakes and accidents, as well as in criminal cases involving homicide, rape, or suicide (5,6). In these situations, forensic odontology often works alongside forensic anthropology to estimate key

characteristics such as age, sex, and race from photographs, bite marks, lip prints, palatal rugae, radiographs, and dental DNA (4).

Bite mark analysis, a key area of forensic odontology, has been widely recognized for its potential evidentiary value. Sweet and Pretty (7) underscored the importance of bite marks as critical forensic evidence, emphasizing that these marks often contain both physical and biological data that can be used to identify perpetrators or exclude suspects. However, accurate analysis requires careful collection and expert interpretation due to the complexity of injuries on elastic and deformable surfaces like human skin, which can complicate investigations in cases of violent crimes. Despite these challenges, this highlights the significant role bite marks play in criminal investigations, given their frequent presence in violent crimes.

Bite mark analysis has come under increasing scrutiny in recent years. While Rawson et al. (8) initially supported the assumption that human dentition is unique, later studies, such as those by Bush et al. (9), challenged this notion. Bush et al. argued that bite mark analysis could be unreliable due to correlations and nonuniformity in tooth arrangements, raising questions about its validity in forensic investigations. More recently, Franco et al. (10) explored the uniqueness of bite marks through a large sample of pairwise comparisons collected via extraoral 3D scanning and Euclidean distance measurements. Their study provided evidence suggesting that human dentition can be unique within large, randomly selected populations. This ongoing debate shows the need for more research to validate assumptions about uniqueness in bite mark analysis.

The NIST Interagency Report (11) emphasized that there was insufficient empirical evidence to support the uniqueness of human dentition as it appeared in bite mark evidence, particularly on human skin. The report also pointed out inconsistencies and subjective interpretations in current bite mark analysis practices, which could undermine the reliability of this evidence in court cases.

Concerns surrounding the reliability of bite mark analysis stem from historical inconsistencies and methodological errors in cases from the 1980s and 1990s, which led to wrongful convictions (12). Technological advancements, such as cone-beam computed tomography (CBCT), have greatly enhanced dental analysis. CBCT enables detailed imaging of both hard and soft tissues, improving the examination of dental features like palatal rugae, tooth occlusion, and soft tissue thickness with greater accuracy (3). Despite these advancements, challenges remain, especially due to the lack of standardized bite mark analysis protocols.

In response to these challenges and criticisms, revised standards and guidelines have been implemented to improve consistency in forensic odontology practice (12). The American Board of Forensic Odontology (ABFO) has introduced independent verification procedures and ongoing certification requirements for practitioners, ensuring that technological advancements align with updated methodological rigor.

Currently, there is no universally accepted method for determining the type, quality, or number of characteristics required for a reliable bite mark comparison. This lack of standardization has led to inconsistent results and has diminished the evidentiary value of bite

marks in legal proceedings. This study aims to bridge these gaps by employing advanced statistical methodologies, such as Principal Component Analysis (PCA), to create a standardized system for comparing bite marks with dental molds. By developing a composite scoring system, this research seeks to establish clear evidentiary thresholds and minimize bias, leading to more reliable and objective forensic analyses.

1.2 Research goal

Forensic odontology plays a crucial role in the identification of individuals based on dental evidence. However, one of its inherent challenges is the lack of a standardized method to determine the evidentiary value of bite mark comparisons. Since bite marks are often subject to individual interpretation and variability, there is no universally accepted protocol to define the type, quality, and number of individual dental characteristics required to conclude that a bite mark provides sufficient evidence for identification. This inconsistency has led to variable results in legal cases, sometimes contributing to wrongful convictions or unreliable evidence, as documented in multiple forensic reviews (13).

In forensic contexts, the primary application of bite mark analysis is to compare a bite mark found on a victim or an object to the dental impressions of a suspected biter. This analysis is crucial when the identity of the biter is unknown and there is a need to establish a link between the suspect and the crime scene or the victim.

The primary goal of this research is to develop a standardized approach for bite mark comparison, addressing the current limitations in forensic odontology. The study focuses on four key objectives: 1) identifying the most

critical dental characteristics for reliable bite mark comparisons; 2) standardizing the precision of measurement processes; 3) establishing clear evidentiary thresholds for legal reliability; and 4) enhancing the overall accuracy and reliability of bite mark-to-individual mapping using statistical techniques.

The essence of this forensic procedure is to determine, from a suspect pool, who might have left the bite mark on a victim in a systematic and accurate way. This involves comparing the bite mark with the dental impressions of various suspects to either exclude or implicate individuals in the crime.

Data Collection Flowchart

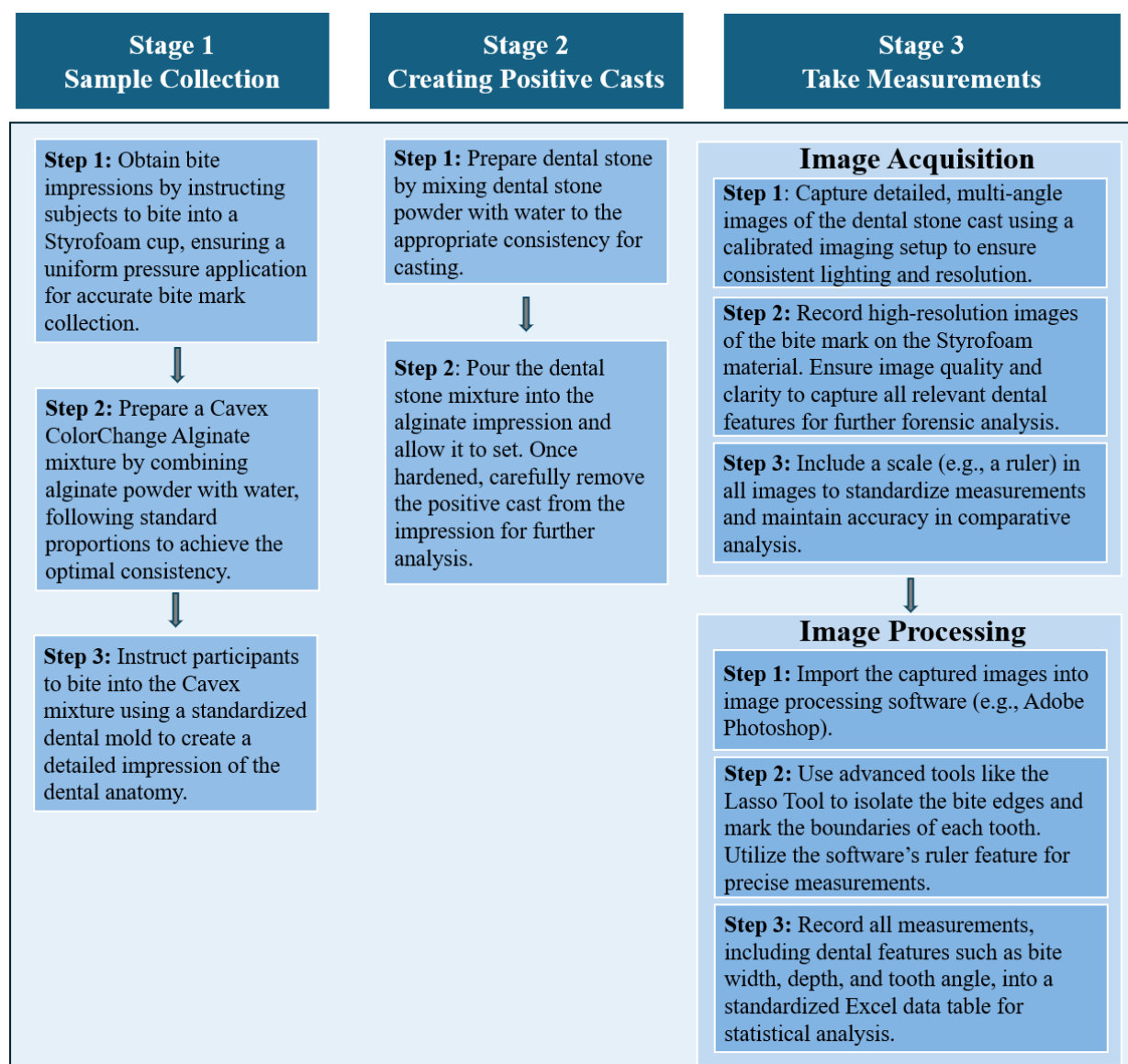


Figure 1. Data Collection Flowchart illustrates the methodology for collecting, processing, and analyzing bite mark evidence, divided into three critical stages: Sample Collection, Creating Positive Casts, and Taking Measurements. Each stage includes detailed steps to ensure accurate and consistent results in forensic odontology.

To achieve these objectives, this study will identify which dental characteristics; such as inter-molar distances, rotation angles, and tooth surface areas; are most reliable for comparing bite marks with 3D molds. It will also establish and recommend threshold values for the number and type of characteristics required to meet a reliable evidentiary standard, ensuring that only highly accurate bite mark comparisons are considered valid in legal contexts and minimizing experimenter bias. Finally, by incorporating statistical techniques such as Principal Component Analysis (PCA), the study aims to develop a composite scoring system that significantly improves the accuracy of bite mark-to-individual mapping. The ultimate aim is to recommend a protocol that forensic odontologists can use in practice, thereby improving the reliability and scientific credibility of bite mark evidence in court.

1.3 Methods

This study's methodology is designed to ensure precise, reproducible results in the collection, processing, and analysis of bite mark evidence. As shown in Figure 1, the methodology follows three critical stages: “sample collection”, “the creation of positive dental casts”, and “image acquisition and measurement”. The combination of high-resolution imaging, advanced software tools, and statistical analysis allows for an objective and standardized evaluation of bite marks and their comparison with dental molds.

2 Data Collection

2.1.1 Materials and tools

The materials and tools utilized were chosen to achieve acceptable precision and accuracy in the collection and analysis of dental impressions. The Cavex ColorChange™ Alginate was selected for its color transition

feature, which indicated the optimal setting time for accurate dental impressions. 250 mL glass beakers were used to ensure complete mixing and measurement of alginate and dental stone, thus maintaining consistency across all samples. Dental trays of various sizes were utilized to accommodate different mouth sizes and to ensure an accurate impression of the teeth. A flat spatula was used for stirring and spreading the alginate and dental stone mixtures, ensuring uniform consistency. A dental vibrator was used to eliminate air bubbles from the dental stone mixture. This device vibrated the mixture gently, allowing trapped air to rise to the surface and ensuring that the final cast remained free of air pockets or defects. This process created a detailed and accurate positive cast of the teeth. A high-resolution camera was employed to capture detailed images of the casts and impressions, ensuring that all relevant dental characteristics were clearly visible for analysis. Adobe Photoshop and Excel were used for image processing and data organization. Photoshop allowed for precise boundary detection of bite marks and dental impressions, while Excel provided a structured way to organize and analyze the resulting data. Python was chosen for its ability to process advanced statistical processing and run Principal Component Analysis (PCA), providing a robust tool for feature selection and dimensionality reduction.

2.1.2 Sample collection – stage 1

The dental samples were collected from 30 high school students, all of whom were 16 years old. Ensuring that all participants were within the same age group was critical for maintaining consistency in the dataset. However, due to the limited sample size, ethnicity and gender sampling were not conducted. While this presented a limitation in terms of demographic diversity, the focus on a consistent age group

provided a controlled environment for analyzing dental characteristics and their relationship to bite mark patterns.

In the first stage of sample collection, negative impressions of participants' teeth were made using Cavex ColorChange™ Alginate. This alginate was selected for its clear color change indicator, which ensured that the impressions were captured at the ideal moment of consistency, reducing the risk of error due to premature or delayed setting. The impressions were carefully handled to avoid distortion, ensuring that the negative molds retained an accurate representation of the individual's dental structure.

During sample collection, first, the alginate was prepared by mixing alginate and water to achieve a consistent, plaster-like texture. Next, participants were instructed to bite into the alginate, ensuring uniform pressure for accurate bite mark creation. Finally, the alginate mixture transitioned from lavender to pink during the setting process, signaling its readiness for removal. Specifically, in a 250 mL glass beaker, 14 grams of Cavex ColorChange™ Alginate powder were measured using two even scoops. The powder was gradually combined with 30 mL of water, stirred with a flat spatula until it reached a plaster-like consistency. As the mixture transitioned from lavender purple to a warm pink, this indicated the material had reached the ideal setting point for capturing the impression. At this stage, the mixed alginate was spread evenly into a dental tray, ensuring no overflow occurred beyond the tray's edges. Participants were instructed to bite down into the alginate-filled tray for ~ 1.5 minutes to ensure a uniform bite mark. Once the alginate cured and turned white, the tray was carefully removed from the participant's mouth and rinsed with cool water to clean the mold. The

impression was left to set for 8-10 minutes before the negative mold was removed from the tray, ensuring it was intact and ready for further use. These steps were repeated for both the upper and lower sets of teeth to ensure that a complete set of impressions was collected for each participant.

This study did not require IRB approval as it was a school-based research project conducted over a limited two-month period, with all participants providing consent. The study did not involve vulnerable populations, sensitive topics, or interventions that posed any risk to participants. Additionally, the dataset contained no personally identifiable information, ensuring both anonymity and confidentiality. Given these factors, the study falls within the scope of research exempt from formal IRB review under standard ethical guidelines for minimal-risk educational research.

2.1.3 Creating positive casts - stage 2

The second stage involved creating positive dental casts from the alginate impressions using dental stone. These casts provided a durable, 3D model of each participant's teeth, which was crucial for accurate comparison with bite mark evidence. The use of a dental vibrator helped eliminate air bubbles, which could have compromised the accuracy of the casts.

The process began by mixing and pouring the dental stone. A precise ratio of dental stone powder to water was used to create a mixture with a creamy consistency, ideal for filling the impressions. After pouring the mixture into the alginate molds, the dental stone was allowed to cure. This curing process ensured that the casts hardened properly, producing an intact, detailed model of the teeth for further analysis. Specifically, in a clean 250 mL glass beaker, 60

Process for Creating Dental Impressions and Casts

Stage 1

Sample Collection

Prepare the mixture and mold to obtain teeth impression from the participants

Making the Negative Impression



Mix Cavex powder & water:
Stir thoroughly till the color change to lavender purple. Lavender purple color indicates that the mixture is ready.

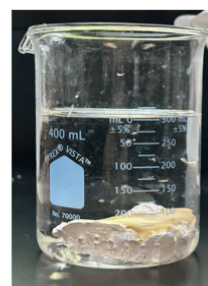


Spread Mixture in the Mold:
Spread the mixture evenly across the mold, avoid air bubbles. The mixture turns warm pink, when it starts to set.

Making the Positive Cast



Making the positive cast:
Pour the dental stone mixture into the negative impression to create the positive mold of the teeth.



Setting the Cast in Water:
Place the cast in water to help set the mixture completely. Leave the cast in water till it hardens properly before handling.

Stage 2

Creating Positive Casts

Making the positive cast using the negative impression.

Finished Positive Cast



Finished positive cast: Once set, remove the positive cast.

Figure 2. Process for Creating Dental Impressions and Casts. These visuals outline the two main stages in the creation of dental impressions and casts: Stage 1 focuses on the preparation of negative impressions using Cavex ColorChange™ Alginate, while Stage 2 details the creation of positive dental casts from these impressions. Each step is shown to demonstrate how accurate and reliable molds are created for forensic analysis.

mL of dental stone powder was combined with 20 mL of water. The mixture was stirred with a flat spatula until it achieved a creamy, peanut butter-like consistency. To remove air bubbles that could compromise the accuracy of the cast, the beaker was gently tapped on a flat surface. For optimal results, a dental vibrator was used

to further eliminate any trapped air in the mixture. The dental stone mixture was carefully poured into the alginate mold, starting from one side of the impression and allowing the mixture to flow steadily toward the opposite side. This ensured that the dental stone filled the mold evenly without overflowing. The dental stone

was allowed to harden for 35-40 minutes, during which time it cooled and set. The mold, now filled with hardened dental stone, was submerged in room-temperature water and soaked for a few minutes to ease the removal of the positive cast. The dental stone cast was gently loosened and removed from the alginate mold, ensuring the cast remained intact and undamaged, ready for further analysis. Figure 2 provides details of the first two stages of the process, demonstrating the preparation of negative impressions in sample collection and the subsequent creation of positive casts in Stage 2.

2.1.4 Image acquisition and measurement – stage 3

Following the creation of the dental casts, the next phase involved detailed image acquisition and measurement to collect the necessary data for subsequent analysis. The process began with capturing high-resolution images of the dental stone casts using a calibrated camera system.

Consistent lighting and the inclusion of a ruler in each image ensured standardized scaling, enabling precise and accurate measurements across all samples.

The captured images were then processed using Adobe Photoshop. The Lasso Tool was employed to isolate the boundaries of each bite mark and tooth impression. By carefully outlining the edges of the dental features, the software's built-in measurement tools were used to calculate key dimensions, including bite width, depth, and tooth angles. These precise measurements allowed for a detailed quantitative assessment of the dental structures. The data points collected were; Inter-molar distance between the first molars (cm), Inter-canine distance between cuspids (cm), Surface area of each tooth's biting edge (cm²), Rotation angle of the first and second bicuspid (°), Tooth surface areas for both right and left molars and bicuspid.

Table 1. List of dental measurements captured for the analysis of the bite marks, including inter-molar distances, tooth surface areas, and rotation angles.

Measurements Collected			
1	Distance Between First Molars	15	Right First Bicuspid Surface Area
2	Distance Between Cuspids	16	Right Cuspid Rotation Angle
3	Distance Between First Bicuspid	17	Right Cuspid Surface Area
4	Right First Molar Rotation Angle	18	Right Lateral Incisor Rotation Angle
5	Right First Molar Surface Area	19	Right Lateral Incisor Surface Area
6	Left First Bicuspid Rotation Angle	20	Right Central Incisor Rotation Angle
7	Left First Bicuspid Surface Area	21	Right Central Incisor Surface Area
8	Left Second Bicuspid Rotation Angle	22	Left Central Incisor Rotation Angle
9	Left Second Bicuspid Surface Area	23	Left Central Incisor Surface Area
10	Left First Molar Rotation Angle	24	Left Lateral Incisor Rotation Angle
11	Left First Molar Surface Area	25	Left Lateral Incisor Surface Area
12	Right Second Bicuspid Surface Area	26	Left Cuspid Rotation Angle
13	Right Second Bicuspid Rotation Angle	27	Left Cuspid Surface Area
14	Right First Bicuspid Rotation Angle		

The key metrics collected for each sample, such as inter-molar distances, tooth surface areas, and rotation angles, are shown in Table 1, providing a structured summary of the measurements utilized for the bite mark analysis.

To calculate the surface area of teeth from the images of dental molds, a pixel-based method

using Adobe Photoshop was employed. The process began by identifying the number of pixels within the tooth's surface area. A known ruler scale (in inches) included in each image was used to establish the pixel-to-inch conversion factor. Equation (1) shows the calculation.

$$\text{Surface Area of Tooth (in cm}^2\text{)} = [(\text{pixels in tooth surface area}) \times (\# \text{ of pixels per inch}) \times (2.54 \text{ cm per inch})]^2 \quad (\text{eq1})$$

This allowed the conversion of pixel counts into real-world measurements in square centimeters. For instance, in a sample calculation, if a tooth's surface area measured 30 pixels and the pixel density was 282 pixels per inch, the resulting surface area would be $\sim 0.73 \text{ cm}^2$.

measurements did not unduly impact the final composite score. By averaging these calculations, the study aimed to produce a more consistent and precise analysis, thereby increasing the robustness of the composite scoring system.

To enhance the accuracy and reliability of the measurements, each sample image was processed 5 times using the software to normalize operator errors. This multi-calculation approach helped reduce variability and ensured that minor inconsistencies in

As illustrated in Figure 3, a dental cast was positioned beside a ruler to provide a reference for scaling, enabling accurate conversion of pixel measurements into real-world dimensions (cm^2) for surface area calculations.



Figure 3: Dental Mold Measurement Calibration: This image illustrates the process of using a ruler to establish a pixel-to-inch conversion factor for calculating the surface area of teeth from digital images.

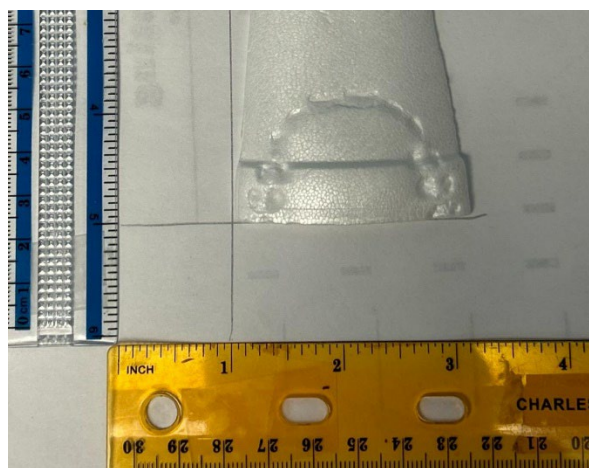


Figure 4. *Measurement Calibration for the bite mark impression on Styrofoam:* This image illustrates the process of using a ruler to establish a pixel-to-inch conversion factor for calculating the surface area of teeth from digital images.

As illustrated in Figure 4, Styrofoam with the bite mark was positioned beside a ruler to provide a reference for scaling, enabling accurate conversion of conversion of pixel measurements into real-world dimensions (cm²) for surface area calculations.

All measurements were systematically recorded in an Excel spreadsheet to ensure proper organization and ease of access for subsequent statistical analysis. Each sample's biting edge surface area, rotation angle, and inter-molar and inter-canine distances were carefully logged to facilitate accurate comparisons between the bite marks and the dental casts.

The Photoshop analysis was conducted over five independent trials, with each sample measured five times to enhance the accuracy and reliability of the results. The measurements from each trial were averaged to ensure consistency and reduce the impact of random errors. Any trial measurement that deviated significantly from the others; defined as being more than two standard deviations from the mean; was excluded as being an outlier. This method of multiple trials and outlier exclusion helped ensure that the final measurements

reflected accurate and consistent data for further analysis.

2.2 Data preparation

Once the measurements were completed and recorded in Excel, the data was prepared for statistical analysis. These measurements served as critical inputs for Principal Component (PCA), which aimed to reduce the dimensionality of the dataset and identify the most relevant features for matching bite marks to dental casts.

To prepare the data, the measurements from each sample were standardized to account for variations in scale and bite strength across different individuals. Standardization involved transforming the data so that each variable was assigned a mean of zero and a standard deviation of one. This step was particularly crucial because PCA is sensitive to the scale of variables. Without standardization, variables with larger variances (such as those measured in larger units) could disproportionately influence the analysis, potentially leading to skewed results. By standardizing the data, each variable contributed equally to the PCA, ensuring that

differences in scale did not affect the identification of principal components.

Additionally, factors such as inter-molar distances, inter-canine distances, and tooth surface areas are likely to be correlated due to their inherent relationship with dental structures. PCA helps to address this collinearity by transforming the original correlated measurements into a set of uncorrelated principal components. These principal components allow for more reliable statistical analysis by reducing redundancy and focusing on independent dimensions of variation.

Finally, the standardized data was imported into Python for statistical analysis and further processing. Here, PCA was applied to determine the principal components that contributed most significantly to the variance in the dataset, allowing for a clearer focus on the key features relevant to bite mark comparison.

2.3 Scoring

Establishing a consistent framework for measuring and evaluating bite mark evidence is crucial in forensic odontology. Variability in measurements and analysis methods can lead to inconsistencies, undermining reliability. To mitigate these issues, this study developed a standardized approach, incorporating statistical techniques such as PCA to improve accuracy and consistency. The detailed scoring methodology of bite mark analysis is shown in Figure 5 which lays out key steps such as the application of PCA, creation of composite scores, and the process of comparing bite marks with dental casts.

2.3.1 Standard Operating Protocol (SOP)

A Standard Operating Protocol (SOP) was developed to ensure consistency in bite mark analysis. This protocol introduced a systematic approach to standardize the comparison of dental impressions with bite evidence. The goal was to provide a clear and reproducible scoring framework that addressed the inherent subjectivity in bite mark analysis. The scoring system assigned a numerical value to each tooth's characteristics based on the measurements taken, thereby promoting consistency and reducing bias in the interpretation of bite mark evidence.

Through the use of Principal Component Analysis, dimensionality reduction was achieved by focusing on key dental features, including inter-molar distance, bite width, and surface area of teeth. These features were then synthesized into a composite score. A composite score for each bite mark and dental cast was created. The formula for the composite score was based on the weighted sum of the top three principal components. This ensured that the model captured the most significant aspects of the bite mark and dental mold, which together explained at least 70% of the variance in the data.

The choice of 70% variance as the cutoff is common in PCA-based analysis because it strikes a balance between reducing dimensionality and retaining enough information for accurate modeling. Capturing 70% of the variance ensures that the majority of the important data patterns are included while minimizing overfitting, which could occur if too many components were used (14).

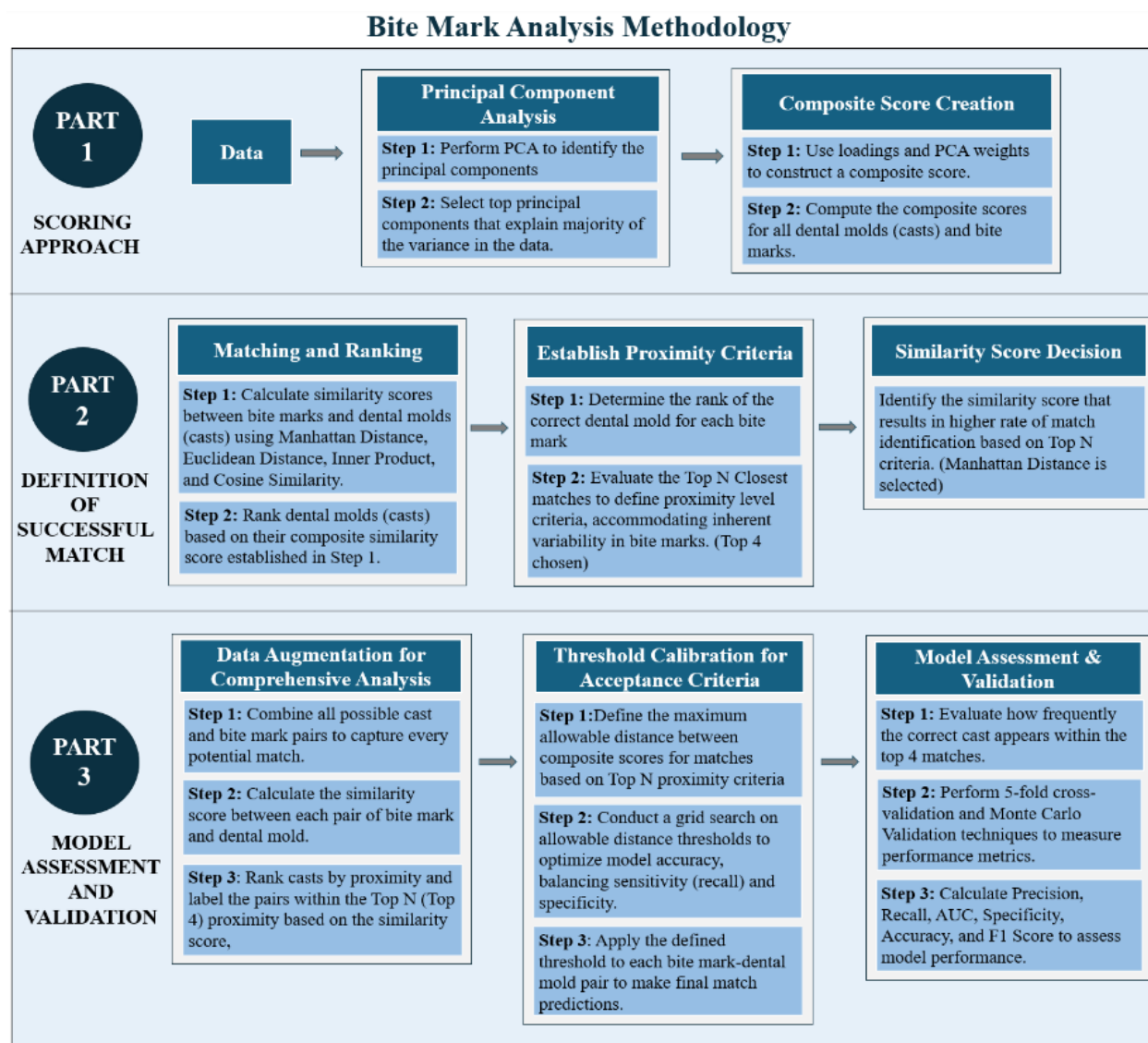


Figure 5. Methodology for Bite Mark Analysis diagram illustrates the step-by-step approach used to create and validate the composite scoring model for bite mark comparisons. It includes key phases like data segmentation, principal component analysis (PCA), composite score computation, threshold establishment, and model validation.

2.3.2 Principal Component Analysis (PCA)

Principal Component Analysis was performed on the measurements obtained from the dental molds. A total of 27 dental measurements were initially calculated for each dental mold sample. To ensure consistency and reliability, five independent trials were conducted for each measurement. However, certain variables were excluded from the final analysis due to large variance observed across these trials, indicating inconsistencies in data collection. Subsequently, PCA was applied to reduce the dimensionality of the data and identify the most

significant variables from the bite mark and dental cast measurements. This statistical technique identified the principal components that contributed most to the variability in the data. By identifying these components, PCA enabled the creation of a more focused and less noisy dataset for analysis. In this study, the top three principal components accounted for 72.08% of the total variance, with the first component (34.15%) representing surface area, the second component (23.52%) representing rotational angles, and the third (14.42%) capturing distances between teeth.

Due to the limited sample size, a modification to the standard methodology was necessary to maximize the robustness of the model without compromising its reliability. Ideally, PCA would be applied to the training set after splitting the data into training and test sets. However, to avoid data fragmentation and to preserve valuable variance patterns, PCA was performed on the entire dataset. This approach ensured that sufficient information was retained in each principal component, creating a comprehensive basis for the composite score calculation.

2.3.3 Composite score development and initial analysis

The top three principal components, which explained 72% of the variance, were selected for inclusion in the composite score. The identified principal components from PCA were used to calculate a composite score for each sample, with the formula based on the weighted sum of the top three components. Each component was weighted according to its explained variance ratio, ensuring that the most significant features contributed proportionally

to the final score for bite mark-to-dental cast comparisons. Measurements that did not provide substantial value in differentiating bite marks were excluded, leaving a final selection of 18 features for the PCA-based framework.

Each principal component is represented by a loading vector of dimension $[1 \times n]$, where 'n' represents the number of features analyzed. For this study, these dimensions are $[1 \times 18]$, as 18 dental measures are examined. As mentioned earlier, the top three principal components are selected based on their ability to explain the maximum variance in the dataset. By selecting the top three components, three loading vectors were identified, each capturing distinct patterns in the data.

To derive the coefficients for the composite score, a weighted average of the loadings across these three principal components was calculated, with weights corresponding to the explained variance of each component. The weighted contribution for each feature i on principal component k for a given data point is calculated as in equation (2)

$$\text{Weighted Contribution}_{i,k} = \text{Explained Variance Ratio}_k \times \text{Loading}_{i,k} \times \text{Standard Value}_i \quad (\text{eq2})$$

where *Explained Variance Ratio_k* is the explained variance ratio for principal component k , *Loading_{i,k}* is the loading of feature i on principal component k , and *Standard Value_i* is the standardized value of feature i for the data point. The weights for the *Explained Variance Ratio_k* were $w_1 = 0.3415$, $w_2 = 0.2352$, and $w_3 = 0.1442$. This process

assigned a unique coefficient to each of the 18 dental features, reflecting their relative contributions to the patterns captured by the principal components. Next, the composite score was calculated by aggregating the weighted contributions of each feature across the top three principal components, as shown in Equation (3).

$$\text{Composite Score} = \sum_{k=1 \text{ to } 3} (\text{Explained Variance Ratio}_k \times \text{Loading}_{i,k} \times \text{Standard Value}_i) \quad (\text{eq3})$$

This procedure resulted in a composite score that integrated the contributions of the top three principal components into an n-dimensional vector, where n=18, representing the selected features. Additionally, non-linearity was addressed by assigning greater weight to the most informative variables in the composite score calculation, enhancing its robustness for forensic applications. This composite score provided a systematic approach to evaluating similarity between bite marks and dental molds, improving the reliability and reproducibility of forensic analyses.

2.3.4 Matching and ranking

The primary objective of the matching and ranking process was to accurately evaluate the similarity between bite marks and dental molds to establish a basis for identifying a "successful match." Given the inherent variability in bite

marks due to factors like pressure, skin elasticity, and anatomical differences, it was essential to determine which similarity measure provided the most reliable results for this application.

To capture different aspects of similarity, we initially tested several similarity scoring methods; Manhattan Distance, Euclidean Distance, Cosine Similarity, and Inner Product; each offering a unique perspective on the relationship between bite marks and molds.

Manhattan Distance, shown in Equation (4), calculates the cumulative absolute difference between each element of the composite score vectors $\vec{X}$ and $\vec{Y}$. Manhattan Distance provides a robust similarity assessment by reducing sensitivity to outliers, making it suitable for handling the variability inherent in bite marks.

$$d_{\text{Manhattan}}(\vec{X}, \vec{Y}) = \sum_{i=1 \text{ to } n} |X_i - Y_i| \quad (\text{eq4})$$

Euclidean Distance, as defined in Equation (5), captures the direct spatial distance between the composite score vectors. This measure is useful

for assessing absolute similarity, as it considers the overall magnitude of difference between vector elements.

$$d_{\text{Euclidean}}(\vec{X}, \vec{Y}) = (\sum_{i=1 \text{ to } n} (X_i - Y_i)^2)^{1/2} \quad (\text{eq5})$$

Shown in Equation (6), Cosine Similarity evaluates the cosine of the angle between score vectors $\vec{X}$ and $\vec{Y}$, which highlights the directional

similarity. This measure is beneficial for identifying pattern-based similarity, as it is independent of scale and focuses on the shape of the vector rather than its magnitude.

$$\text{Cosine Similarity}(\vec{X}, \vec{Y}) = \vec{X} \cdot \vec{Y} / \|\vec{X}\| \|\vec{Y}\| \quad (\text{eq6})$$

Defined in Equation (7), the Inner Product emphasizes the angular alignment between the vectors, capturing how well the directional

aspects of the scores match. This approach is effective in identifying linear relationships between the features of bite marks and molds.

$$\text{Inner Product}(\vec{X}, \vec{Y}) = \sum_{i=1 \text{ to } n} X_i \cdot Y_i \quad (\text{eq7})$$

This systematic approach allowed for a rigorous comparison of the effectiveness of each similarity metric in capturing the underlying relationship between bite marks and molds, as it provided robust performance in accounting for overall similarity while minimizing sensitivity to outliers. Selecting the best-performing metric enhanced the accuracy of the matching process, ensuring that only the most suitable measure of similarity was applied in further analysis and threshold calibration. In this analysis, all the attributes were first standardized before using in the similarity calculations. For this purpose, z-score normalization was used. The distance measures were calculated on the standardized data so that the analysis was not biased by the inherent differences in attribute scales and units.

2.3.5 *Establishing the proximity criteria (success criteria)*

The primary objective was to determine the degree of similarity between bite marks and dental molds. First, similarity scores were calculated using Manhattan Distance, Euclidean Distance, Inner Product, and Cosine Similarity for the cast and bite mark pairs of each subject in the study.

A critical step in this research involved defining what constituted a "successful match" between a bite mark and a dental mold. Given that bite marks can be imperfect and inherently variable, it was insufficient to rely solely on the closest match based on composite scores. To address this, an initial analysis was conducted to assess various "proximity criteria." Rather than considering only the single closest match, the analysis was expanded to evaluate the top 2 to top 10 closest matches based on differences in composite scores. This multi-tiered approach provided insights by evaluating matches at multiple ranked levels. The process involved calculating the difference in composite scores

between each bite sample and every dental mold, then ranking the molds by their proximity to each bite's composite score. This method allowed for capturing subtle variations in bite mark comparisons, providing a more robust framework for determining successful matches. After evaluation, Manhattan Distance was selected as the most reliable metric for this application, and the Top 4 closest matches were chosen as the success criteria.

2.3.6 *Data augmentation for comprehensive analysis*

In forensic odontology, the process of matching bitemarks with dental casts mirrors real-world forensic investigations where multiple possible pairings must be evaluated to identify the correct match. To simulate this scenario, an exhaustive pairwise matching approach was employed to generate a dataset that included every possible pairing of bitemarks and casts. This ensured that each cast was compared with each bitemark, forming an augmented dataset of systematic cross-pairings beyond the original matched pairs.

This approach was particularly valuable for assessing the model's robustness and specificity in distinguishing true matches from non-matches under realistic conditions. By including both matching and non-matching pairs, the model could be evaluated on its ability to accurately identify the true match among several candidates—a critical aspect in forensic contexts. The key benefits of this augmented dataset included *Real-World Simulation*: In practical forensic investigations, a bitemark may need to be compared against multiple dental casts. This augmented dataset mirrored that process, allowing us to evaluate how well the model ranked potential matches when presented with several candidates. *Accurate Reflection of Performance Metrics*: By

including both matching and non-matching pairs, the augmented dataset provided a more realistic basis for calculating performance metrics such as precision, recall, specificity, and F1 score. This setup ensured that these metrics accurately reflected the model's capability to identify true matches and exclude false positives, making the evaluation more representative of forensic scenarios. *Optimized Thresholds for Matching:* Testing a range of cast-to-bitemark distances within the augmented dataset facilitated the fine-tuning of similarity thresholds. By optimizing the distance threshold for a "successful match," the model's reliability in correctly identifying close matches was enhanced. *Ranking Capability Assessment:* This approach allowed for the evaluation of how often the correct match ranked within the top few candidates, a scenario that mirrors real life forensic applications where multiple close matches may need to be shortlisted.

In summary, this systematic cross-pairing approach provided a rigorous and realistic testing framework, enabling the model to be evaluated as it would be in a real-world forensic investigation, where identifying the correct match among multiple candidates is critical.

2.3.7 Threshold calibration for the acceptance criteria

Following the establishment of a multi-tiered matching framework, the next step was to define a precise and reliable threshold for match acceptance, a process known as threshold calibration. This calibration aimed to set the maximum allowable deviation between the composite scores of bite marks and corresponding dental molds, ensuring that only the most similar pairs were accepted as matches, as represented in equation (8). The goal was to ensure that the chosen threshold provided a robust and reliable basis for comparison, effectively minimizing both false positives and false negatives.

$$|(\text{Composite Score}_{\text{Bite}} - \text{Composite Score}_{\text{Mold}})| \leq \Delta \quad (\text{eq8})$$

where $\text{Composite Score}_{\text{Bite}}$ is the composition score of the bite mark, $\text{Composite Score}_{\text{Mold}}$ is the composition score of the dental mark, and Δ is maximum acceptable difference. A range of possible threshold values for Δ were tested through a grid search to identify the optimal threshold that minimized both false positives and false negatives. The final threshold was chosen based on its ability to consistently align predicted matches with the success criteria established by the Top 4 proximity criteria, providing a robust and reliable basis for comparison across varying bite mark patterns and dental molds.

2.3.8 Model performance and validation

The augmented dataset used for this analysis originated from an initial set of 30 full matches, making it essential to maximize the information available for training and evaluation. To make the most of the limited data, a combination of 5-fold cross-validation and Monte Carlo cross-validation was employed, offering complementary insights into model performance and robustness. This approach allowed the model to be tested across a range of scenarios, minimizing the impact of data variability on model assessment and enhancing the generalizability of results.

5-fold cross-validation served as the primary validation method. Here, the augmented dataset was divided into five equal segments, or "folds." Each fold was treated as the test subset in turn, comprising 20% of the data, while the model's predictions on these test subsets were evaluated. This rotation through each fold ensured that every observation contributed to the evaluation, providing a balanced assessment of performance metrics across all folds. This approach allowed for a thorough evaluation of the model's stability, reliability, and predictive accuracy across different subsets of the data.

To further validate the model's robustness beyond fixed folds, Monte Carlo cross-validation was also applied. In this approach, the dataset was repeatedly shuffled and randomly split into subsets, with 20% designated as test data in each iteration. The model's predictions were assessed on these randomized subsets, with the process repeated multiple times with different random seeds. Monte Carlo cross-validation provided insights into the model's performance across diverse random configurations, enhancing understanding of its resilience to data variations.

Both validation methods provided a comprehensive set of performance metrics, including Recall, Precision, F1 Score, Accuracy, Specificity, PR-AUC, and AUC. The inclusion of Monte Carlo validation, in addition to 5-fold cross-validation, allowed for a more nuanced understanding of the model's predictive power across varied data shuffles, further supporting the robustness of the selected threshold and similarity measure. By combining these two validation techniques, the study achieved an in-depth perspective on the model's stability, reliability, and generalizability, underscoring the confidence in the results obtained.

2.3.9 Addressing dataset imbalance

In forensic odontology, datasets often exhibit inherent imbalance due to the rarity of true matches compared to non-matches. This natural imbalance was further amplified in this study through the use of data augmentation, which generated an exhaustive set of pairwise comparisons between bite marks and dental molds. While this approach enhanced the dataset's comprehensiveness, it also increased the proportion of non-matching pairs, reflecting real-world scenarios where potential suspects far outnumber actual matches.

Imbalances can distort traditional performance metrics like recall, accuracy, AUC, and specificity, potentially exaggerating the model's success in identifying non-matches while undervaluing its ability to detect true matches. To address these challenges and provide a more balanced evaluation of the model's performance, additional metrics including Precision-Recall Area Under the Curve (PR-AUC), F1, and F2 scores were calculated. Unlike traditional ROC AUC, PR-AUC is specifically designed to evaluate the model's performance in distinguishing the minority (positive) class in imbalanced datasets. PR-AUC focuses on precision and recall, offering a clearer representation of the model's ability to accurately identify true matches while avoiding false positives.

The F1 score provides a balanced assessment of the trade-off between precision and recall, treating both metrics with equal importance. It is particularly useful when both false positives and false negatives carry significant consequences. The F2 score, on the other hand, prioritizes recall over precision by assigning greater weight to false negatives. This makes it especially relevant for forensic applications,

where failing to identify a true match can have severe implications.

By incorporating PR-AUC, F1, and F2 metrics alongside traditional measures, this study

ensures a comprehensive and robust evaluation of the model's capability to handle the challenges posed by dataset imbalance, ultimately enhancing its practical utility in real-world forensic scenarios.

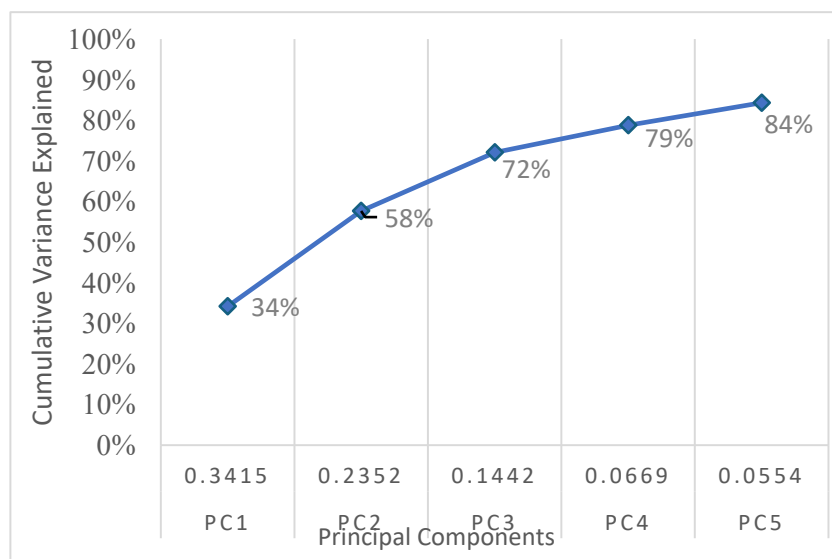


Figure 6. Principal Component Analysis Results. The figure illustrates the variance explained by each of the top five principal components, highlighting how the top three components account for over 72% of the total variance in the dataset. These components were used to generate feature-level weights for bite mark-to-dental cast comparisons.

3 Results

3.1 Standardized scores

The Principal Component Analysis (PCA) found that the top three principal components accounted for 72.08% of the total variance in the data. Specifically, the first principal component (PC1) explained 34.15% of the variance and was primarily associated with the overall surface area of the teeth. This suggested that differences in tooth surface areas played a dominant role in distinguishing individual dental patterns. The second principal component (PC2) explained 23.52% of the variance and captured the angle of rotation of the teeth, indicating that rotational positioning of the teeth significantly contributed to the variability in bite marks. Finally, the third principal component (PC3) explained 14.42%

of the variance and was associated with inter-molar and inter-canine distances. Figure 6 depicts the variance explained by each principal component.

Utilizing the identified principal components, a composite scoring methodology was developed to rank dental casts relative to each bite mark based on their statistical similarity to ensure accurate matching. The contributions of individual dental features to the composite score were quantified by weighting each feature according to the variance explained by its respective principal component. This approach enabled a shift from a generalized PCA-based framework to a focused, feature-specific scoring system. By applying this method, the model assigned coefficients to each relevant dental feature, prioritizing those with the most

statistical significance in the scoring process. This not only enhanced the accuracy of the model but also reduced noise from less informative features. Table 2 presents the coefficients for the 18 dental measures that were found to be significant contributors to the composite score, emphasizing features with a crucial impact on the matching process.

Table 2. Composite score features and corresponding coefficients: This table presents the final coefficients for the 18 dental measurements that contributed significantly to the composite scoring system. Each coefficient is derived from the variance explained by the respective principal component, providing a detailed feature-level scoring system for bite mark analysis. The coefficients are scaled by a factor of 100 to convert them into percentage-like values, enhancing readability and interpretability by presenting the contributions of each dental feature in a more familiar range.

Composite Score Features and Coefficients		
Label	Dental Measure	Coefficient (Scaled x100)
A	First Inter-Molar Distance Between First Molars (cm)	13.4
B	Inter-Canine Distance Between Cuspids (cm)	14.0
C	Second Inter-Molar Distance Between First Bicuspid (cm)	14.6
D	Right First Molar Angle of Rotation (in degrees)	0.4
E	Right First Molar Biting Edge Surface Area (cm ²)	8.0
F	Left First Bicuspid Biting Edge Surface Area (cm ²)	14.1
G	Left Second Bicuspid Biting Edge Surface Area (cm ²)	16.0
H	Left First Molar Biting Edge Surface Area (cm ²)	10.3
I	Right Second Bicuspid Biting Edge Surface Area (cm ²)	16.0
J	Right Second Bicuspid Angle of Rotation (in degrees)	1.7
K	Right First Bicuspid Angle of Rotation (in degrees)	-0.3
L	Right First Bicuspid Biting Edge Surface Area (cm ²)	16.6
M	Right Cuspid Biting Edge Surface Area (cm ²)	-6.4
N	Right Lateral Incisor Biting Edge Surface Area (cm ²)	-1.3
O	Right Central Incisor Biting Edge Surface Area (cm ²)	8.8
P	Left Central Incisor Biting Edge Surface Area (cm ²)	1.6
Q	Left Lateral Incisor Biting Edge Surface Area (cm ²)	-2.9
R	Left Cuspid Biting Edge Surface Area (cm ²)	-6.9

Following the assignment of coefficients to each dental measure, the final composite score equation was achieved which is presented in equation (10).

$$\text{Composite Score Vector} = [13.4A, 14.0B, 14.6C, 0.4D, 8.0E, 14.1F, 16.0G, 10.3H, 16.0I, 1.7J, -0.3K, 16.6L, -6.4M, -1.3N, 8.8O, 1.6P, -2.9Q, -6.9R] \quad (\text{eq10})$$

In this equation, each label (A through R) represents a distinct dental measure, as shown in Table 2, with the associated coefficients indicating the weighted importance of each feature as derived from the principal component analysis. These composite scores were computed for both the bite marks and the dental molds, allowing the difference between the composite score of each bite mark and the

respective mold to be analyzed for accuracy in mold-to-bite-mark matching.

3.2 Acceptance criteria for a successful match

To quantify the similarity between bite samples and molds, various distance and similarity measures were applied, each offering a unique

perspective on the relationship between composite score vectors: Manhattan Distance, Euclidean Distance, Cosine Similarity, and Inner Product. Table 3 shows the cumulative match rates for bite samples to dental molds using Euclidian Distance, Manhattan Distance, Inner Product and Cosine Similarity.

Table 3. This table shows the cumulative match rates for bite samples matched to dental molds across various rank thresholds by different similarity measures. For each proximity rank level (e.g., Top 1, Top 2), the table displays both the number of matches (#) and the corresponding cumulative percentage (%) of correct matches.

Cumulative Match Rates by Proximity Rank and Similarity Measure								
Proximity Rank Top N	Euclidian Distance		Manhattan Distance		Inner Product		Cosine Similarity	
	#	%	#	%	#	%	#	%
Top 1	9	30%	7	23%	7	23%	8	27%
Top 2	10	33%	11	37%	9	30%	10	33%
Top 3	16	53%	17	57%	12	40%	13	43%
Top 4	18	60%	21	70%	16	53%	17	57%
Top 5	18	60%	22	73%	18	60%	19	63%
Top 6	19	63%	23	77%	22	73%	22	73%
Top 7	21	70%	24	80%	23	77%	22	73%
Top 8	21	70%	26	87%	25	83%	25	83%
Top 9	22	73%	27	90%	26	87%	25	83%
Top 10	22	73%	27	90%	26	87%	26	87%

As shown in Table 3, focusing solely on the top 1 closest mold yielded match rates ranging from 23% to 30% across different metrics. However, expanding the criteria to include the top 3 closest molds improved accuracy, with match rates increasing to ~ 40% to 57%. Further expanding the criteria to include the top 4 closest molds resulted in match rates of 53% to 70% for different similarity metrics. This finding underscores the importance of considering multiple potential matches, as the

closest match based on composite scores may not always be correct due to inherent variability in bite marks.

Figure 7 provides a visual representation of the cumulative match rates achieved using different similarity measures across various proximity ranks, facilitating a comparison of each measure's effectiveness in identifying correct matches within the closest-ranked molds.

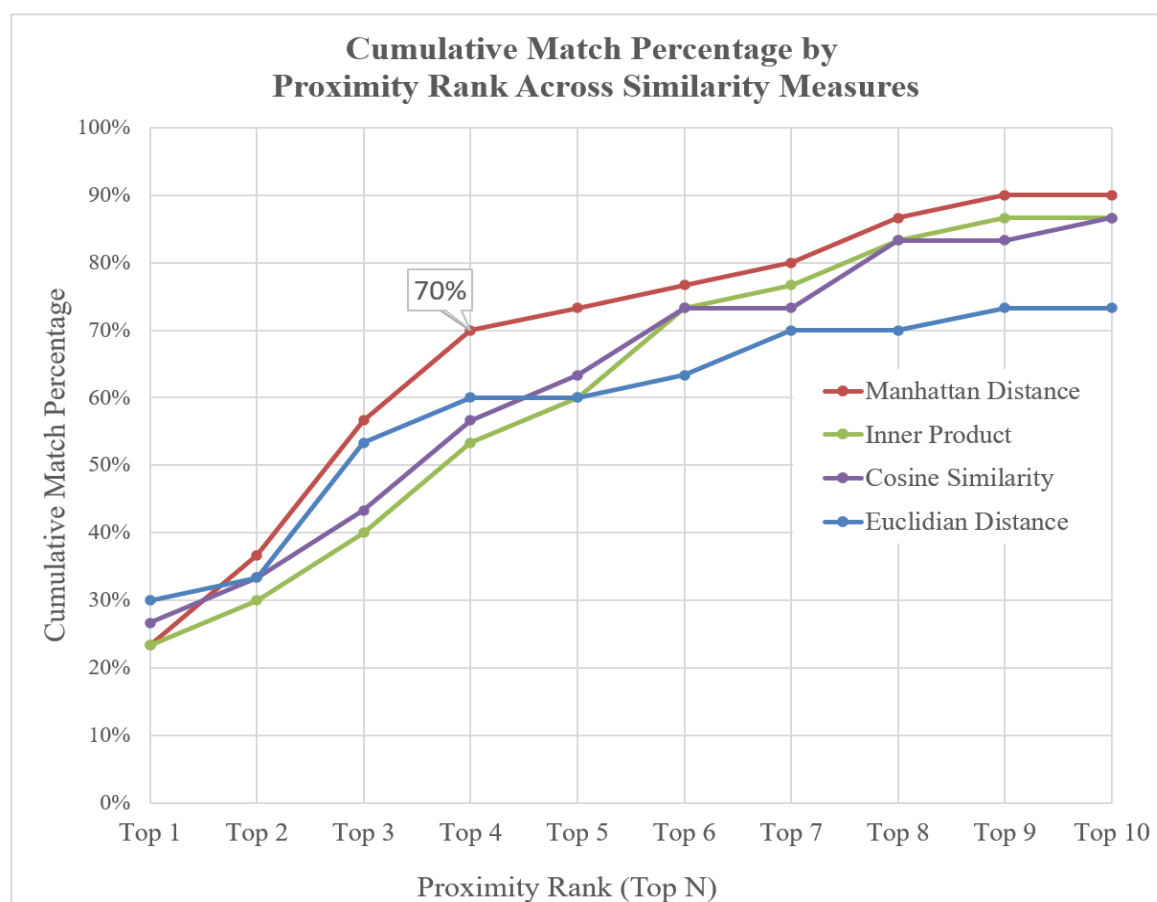


Figure 7. Cumulative Match Percentage by Proximity Rank Across Similarity Measures. This graph displays the cumulative percentage of correct matches for bite samples to dental molds at each proximity rank level (Top 1 through Top 10) across four similarity measures: Manhattan Distance, Euclidean Distance, Inner Product, and Cosine Similarity. The Manhattan Distance approach reaches a cumulative match rate of 70% at the Top 4 rank, outperforming other measures, which require ranks of Top 6 or higher to achieve comparable match rates. This highlights the effectiveness of Manhattan Distance in identifying accurate matches within fewer proximity ranks, supporting its selection as the preferred metric for this analysis.

Based on the cumulative match rates summarized in Table 3 and illustrated in Figure 6, the similarity approach using Manhattan Distance was found to capture more matches within the top 4 ranks compared to the other similarity measures. Additionally, other similarity approaches required ranks of Top 6 or Top 7 to achieve cumulative match rates of 70% or higher, further supporting the selection of Manhattan Distance as the most efficient and effective criterion.

when expanding the proximity ranks beyond this level. Specifically, with Manhattan Distance, expanding the proximity rank from Top 3 to Top 4 resulted in a significant increase in the match rate, increasing from 57% to 70%. However, further expanding the proximity rank from Top 4 to Top 5 only marginally increased the match rate by 3 percentage points, from 70% to 73%. Given these diminishing returns beyond the Top 4, the Top 4 rank was selected as the success criterion.

The decision to focus on the Top 4 ranks was informed by the diminishing returns observed

This decision aligned with practical considerations in forensic investigations. By

including the top 4 closest matches, the likelihood of identifying the correct match increases without excessively broadening the suspect pool, thus reducing the risk of overlooking potential true matches. In real-world scenarios, investigators typically have a much smaller suspect pool than the sample size used in this study. With fewer suspects to compare, the probability of correctly identifying a match within the top 4 closest molds is likely higher, as there is less chance of proximity-based matches among unrelated samples.

Figure 8 depicts the distribution of Manhattan distances across all possible dental cast–bite mark pairings in the augmented dataset. This systematic pairing includes both true matches and non-matches, resulting in a variability of similarity scores that reflects the challenge of

distinguishing true matches in forensic contexts. The left-skewed distribution indicates that most pairings fall within a moderate range of similarity, with fewer pairings exhibiting extreme dissimilarity or high similarity.

This variability was a desired feature of the dataset, as it enabled the evaluation of similarity measures and supported the selection of an optimal threshold for match determination, without introducing bias into the analysis. The distribution also shows the importance of threshold calibration to balance the trade-off between false positives and false negatives, ensuring robust performance across diverse scenarios. The peak in the middle range reflects the realistic complexity of distinguishing true matches from non-matches within a forensic investigation framework.

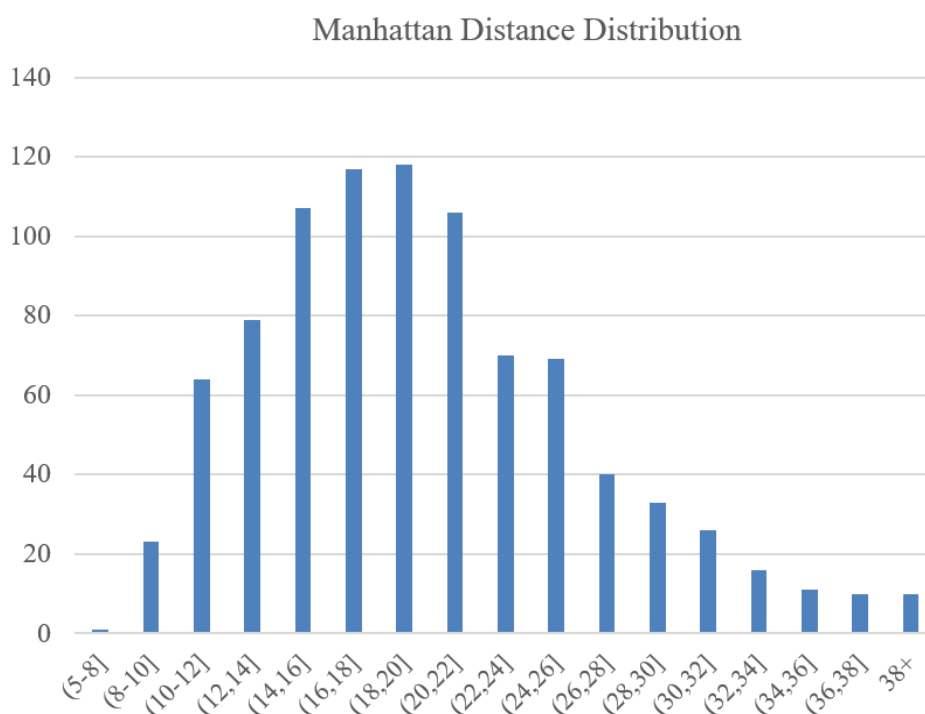


Figure 8. Manhattan Distance Distribution for Augmented Dental Cast–Bite Mark Pairings. This histogram illustrates the distribution of Manhattan distances for a systematically generated Cartesian product of all possible dental mold and bite mark pairings. The majority of pairings exhibit moderate dissimilarity, reflecting the nature of the augmented dataset, which includes both matching and non-matching pairs. The distribution highlights the variability in similarity scores and demonstrates the importance of precise threshold calibration to effectively distinguish true matches from non-matches.

To determine an optimal threshold for identifying a match, a grid search was conducted to systematically evaluate various acceptance thresholds. The goal was to find the threshold that maximized key performance metrics such as accuracy, recall, precision, and F1-score. The acceptance threshold was defined based on the difference between the composite

scores of the bite mark and the dental mold, as shown in equation (8). The grid search identified a threshold of 12.65 points as the best trade-off between minimizing false positives and false negative. The acceptance criteria for defining a match between a bite mark and a dental mold is represented by equation (11).

$$\text{Acceptance Criteria} = |\text{Composite Score}_{\text{Bite}} - \text{Composite Score}_{\text{Mold}}| \leq 12.65 \quad (\text{eq11})$$

Figure 9 provides a comprehensive analysis of how performance metrics varied across different thresholds, enabling an informed selection of the optimal acceptance criterion. Each curve represents a specific metric, such as specificity, recall, precision, F1 score, F2 score, accuracy, ROC AUC, and PR AUC, as a function of the threshold. The dashed line marks the selected threshold of 12.65, where key metrics converge to achieve the best trade-off between false positives and false negatives.

The analysis of assessment metrics across varying thresholds highlights the critical trade-offs in model performance. Specificity decreases and recall increases as the threshold rises, reflecting the shift from minimizing false positives to prioritizing true match detection. Precision and recall intersect near the optimal threshold of 12.65, where F1 and F2 scores also peak, reinforcing this threshold as the most balanced point between precision and recall.

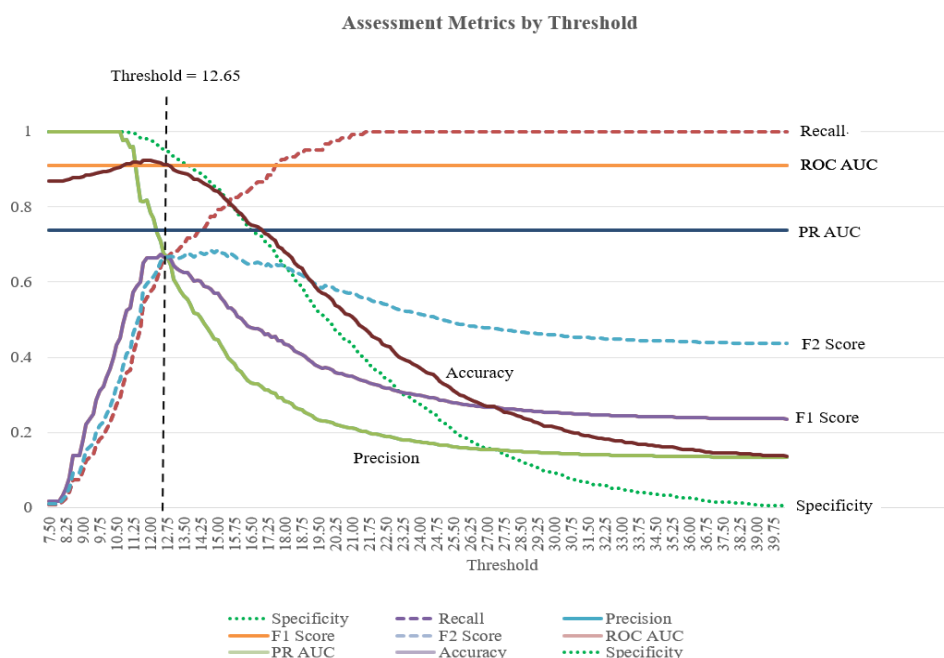


Figure 9. Assessment Metrics by Threshold. This graph displays the variation of key performance metrics, including specificity, recall, precision, F1 score, F2 score, accuracy, ROC AUC, and PR AUC, across a range of thresholds. The dashed line at a threshold of 12.65 marks the optimal point, achieving a balance between minimizing false positives and false negatives while maximizing overall model performance.

At lower thresholds, the model demonstrates high precision but risks missing true matches (low recall), while higher thresholds improve recall at the expense of precision. Meanwhile, AUC values (both ROC-AUC and PR-AUC) remain constant, as they summarize overall model performance across all thresholds. The threshold of 12.65 emerged as an optimal choice, effectively balancing false positives and false negatives to ensure robust performance for forensic bite mark analysis. This calibrated threshold supports accurate and reliable decision-making in real-world forensic scenarios. This tiered ranking approach enhances the precision of the bite mark identification process by balancing the model's accuracy with its practical applicability in distinguishing between very similar molds. Expanding the success criteria to include the top 4 closest molds is not only a practical

compromise but also reflects the realities of forensic investigations, where identifying the correct match from a few close candidates is often the most realistic and effective goal.

3.3 Performance and validation of the proposed composite score definition

The performance of the scoring methodology was evaluated based on several key metrics, including accuracy, precision, recall, F1 score, AUC (Area Under the Curve) and specificity. These metrics were chosen because they collectively indicate the model's effectiveness in correctly identifying matches, minimizing errors, and maintaining reliability. Accuracy measures the proportion of true predictions (both true positives and true negatives) among all cases examined, providing an overall view of model performance.

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{Total Predictions}) \quad (\text{eq12})$$

Precision reflects the proportion of true positive predictions out of all positive predictions made. This metric is vital in minimizing false positives, which could incorrectly widen the

suspect pool in forensic cases. High precision ensures that identified matches are genuinely reliable.

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \quad (\text{eq13})$$

Recall measures the model's ability to correctly identify all actual matches. This metric focuses on minimizing false negatives, crucial in

forensic scenarios where missing a match could allow a perpetrator to go free.

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (\text{eq14})$$

F1 score is the harmonic mean of precision and recall, providing a balanced assessment when

both false positives and false negatives are considered critical.

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (\text{eq15})$$

F2 score is a weighted harmonic mean of precision and recall that places greater emphasis on recall, making it particularly useful in scenarios where false negatives are more critical than false positives. The F2 score assigns four times more weight to recall than to precision.

$$F2\ Score = 5 \times (Precision \times Recall) / (4 \times Precision + Recall) \quad (eq16)$$

Specificity reflects the proportion of correctly predicted true negatives among all actual negative cases. It emphasizes minimizing false positives and becomes particularly significant when wrongly accusing an innocent individual is a concern.

$$Specificity = True\ Negatives / (True\ Negatives + False\ Positives) \quad (eq17)$$

AUC (Area Under the Curve) reflects the model's ability to distinguish between classes across all threshold levels, summarizing performance in a single metric. PR-AUC (Precision-Recall Area Under the Curve) is particularly useful for imbalanced datasets. It focuses on the minority (positive) class performance by emphasizing the balance between precision and recall.

In this analysis, both 5-fold cross-validation and Monte Carlo cross-validation were employed to assess the generalizability and robustness of the threshold applied to the Manhattan Distance metric for matching bite samples to dental molds. Unlike traditional machine learning applications, where cross-validation is often used for model training, here it serves as a tool for performance evaluation and threshold fine-tuning. Given the relatively small sample size and the need to retain all observations, a separate hold-out set was not created. Instead, cross-validation enabled evaluation of the chosen threshold across different subsets of the data, ensuring consistent and reliable performance without sacrificing valuable data for testing purposes.

The performance metrics presented in Table 4 show the efficacy of the proposed composite score methodology in identifying matches

between dental molds and bite marks. Successfully matching bite marks to dental molds is crucial in forensic investigations for identifying potential suspects or clearing innocent individuals of suspicion.

As presented in Table 4, the Cross Validation Results across 5-fold Validation and Monte Carlo Validation were very close to each other, highlighting the model's consistency. In this study, Monte Carlo Validation was performed using 100 random sub samples, which increased its generalizability. As a result, Monte Carlo Validation results will be referenced in the paper as each of these metrics is discussed in this section.

Among the evaluated metrics, PR-AUC, F1, and F2 scores stand out as the most critical for assessing model performance due to the inherently imbalanced nature of the dataset. PR-AUC, which focuses on the balance between precision and recall, is particularly suited to scenarios where the minority class (true matches) holds greater significance. With a PR-AUC of 0.75 in Monte Carlo validation, the model demonstrated its ability to effectively identify true matches while minimizing false positives—a vital requirement in forensic applications.

Table 4. Performance Metrics based on 5-Fold Validation and Monte Carlo Validation: This table presents the performance of the proposed composite score in identifying a bite-mark to mold match.

Model Performance		
Cross Validation Results		
	5-Fold Validation	Monte Carlo Validation
Accuracy	0.91	0.91
Precision	0.69	0.67
Recall	0.69	0.69
F1 Score	0.68	0.68
F2 Score	0.68	0.69
Specificity	0.95	0.95
AUC (Area under the ROC Curve)	0.91	0.91
PR-AUC (Area under the Precision-Recall Curve)	0.74	0.75

The F1 and F2 scores of ~ 0.68 and 0.69 , respectively, indicated a reasonable level of performance for a forensic application involving imbalanced data. These metrics reflect the model's ability to achieve a moderate balance between precision and recall, with the F2 score slightly favoring recall to prioritize the identification of true matches. Given the inherent challenges of forensic bite mark analysis, these scores are consistent with what can be expected in real-world scenarios where class imbalance and measurement variability complicate predictive accuracy.

The overall accuracy of the model was 0.91 , indicating that 91% of the model's predictions (both matches and non-matches) align with the ground truth, defined by successful identification within the Top 4 matches. While this high accuracy demonstrates the model's general reliability, accuracy alone can be misleading in forensic applications, particularly when dealing with imbalanced datasets. Accuracy does not distinguish between the model's performance on the majority (non-matching) and minority (matching) classes, and

it may obscure potential weaknesses in identifying true matches.

With a precision of 0.67 , the model was correct 67% of the time when it predicted a match. Precision is crucial in forensic contexts, where limiting false positives is essential to avoid wrongly implicating innocent individuals. This level of precision suggests that, while the model was relatively effective, a modest rate of false positive identifications remained. However, it is important to note that bite mark analysis is typically one component in a broader body of evidence, rather than the sole factor in convicting an individual. As such, this precision level supports the model's value as a complementary tool, assisting investigators without carrying undue risk of misidentification when considered alongside other forensic evidence.

The recall of the model was 0.69 , showing that it correctly identified 69% of all actual matches. High recall is essential in forensic applications to avoid false negatives, ensuring that true matches are not overlooked. A recall of 0.69

suggested that while the model was effective in identifying true matches, there is room for improvement to further reduce false negatives. The model's specificity of 0.95 indicated its strong ability to correctly identify non-matches, as it correctly rejected 95% of cases that are not true matches. High specificity is especially valuable in forensic analysis to avoid implicating innocent individuals, as it ensures that non-matching dental molds are reliably excluded from consideration. Although the model's precision was lower at 0.67, meaning that some predicted matches were incorrect, the high specificity reflected a strong ability to filter out non-matching samples accurately, reducing the risk of wrongful association. Lastly, the AUC (Area Under the Curve) of 0.91 reflects the model's acceptable discriminative power. A high AUC score shows that the model can effectively distinguish between matches and non-matches across a range of threshold values, making it a robust tool for forensic applications. The AUC score of 0.91 demonstrated the model's ability to maintain high performance across varying decision thresholds, enhancing confidence in its applicability to diverse forensic cases. However, it is important to note that AUC, like accuracy, is not sensitive to the underlying class imbalance, making it a less reliable metric in this context.

While these metrics collectively provided a well-rounded evaluation of the model's performance, they should be interpreted in the context of forensic applications where data imbalance and measurement variability are inherent challenges. Metrics such as PR-AUC, F1, and F2 scores, which are more sensitive to the minority class, are emphasized as primary indicators of the model's reliability in identifying true matches without introducing significant bias or error. The evaluation of the model's performance is shown in Figure 10,

which displays the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the full dataset, Monte Carlo validation, and 5-fold cross-validation. These curves illustrate the model's ability to discriminate between matches and non-matches across varying thresholds, with the corresponding Area Under the Curve (AUC) values providing a quantitative measure of performance. The ROC Curve in Figure 10 demonstrates that the model achieves consistently high AUC values across all evaluations: 0.91 for both Monte Carlo and 5-fold cross-validation, and 0.91 for the full dataset. This indicates strong discriminatory power, suggesting that the model effectively separates matches from non-matches. The PR Curve highlights the model's ability to balance precision and recall, particularly in handling imbalanced data. The PR-AUC values range from 0.74 (5-fold cross-validation) to 0.75 (Monte Carlo validation), further supporting the model's robustness in capturing true matches while minimizing false positives. These values are consistent with the PR-AUC obtained from the full dataset (0.74), underscoring the generalizability of the model's performance.

The performance metrics for the model, such as precision, recall, F1 score, F2 score, and accuracy, are influenced by the choice of threshold for the Manhattan Distance similarity measure. This threshold determines the cut-off point at which a predicted match is classified as positive, allowing investigators to optimize the model's performance to suit specific forensic goals. For example, when avoiding false positives is critical to prevent implicating innocent individuals, a lower threshold prioritizing precision can be used. Conversely, when it is essential to minimize missed matches, a higher threshold maximizes recall, though at the expense of increased false positives.

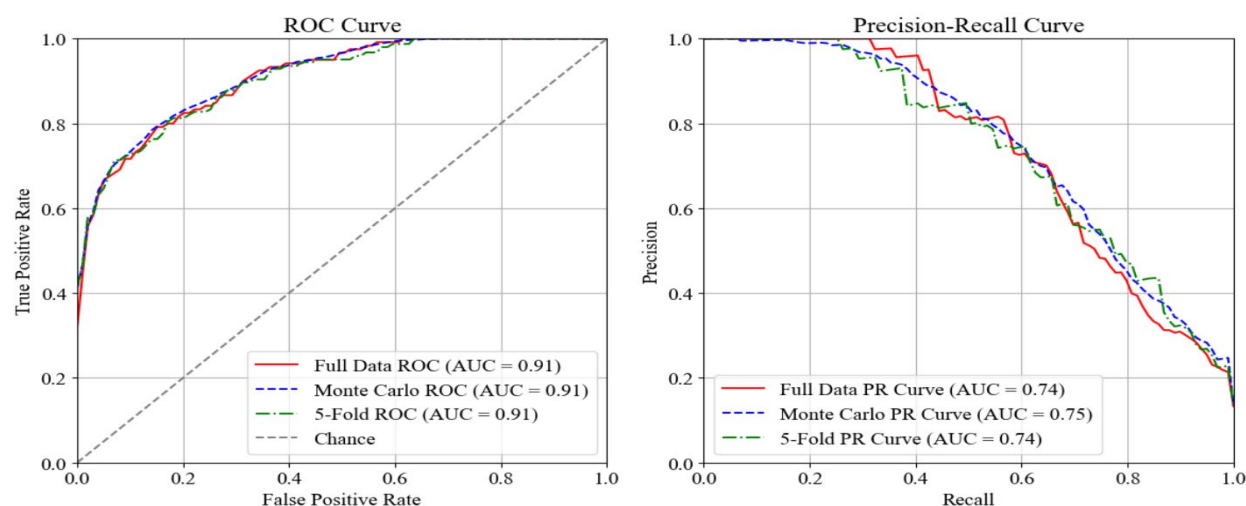


Figure 10. Comparison of ROC and Precision-Recall (PR) Curves for Full Data, Monte Carlo Validation, and 5-Fold Cross-Validation. The ROC curve (left) illustrates the model's discriminative ability, with the Full Data, Monte Carlo, and 5-Fold curves achieving consistent AUC values of 0.91, indicating strong performance in distinguishing matches from non-matches. The PR curve (right) highlights the balance between precision and recall, particularly relevant for imbalanced datasets. The PR-AUC values of 0.74 (Full Data), 0.75 (Monte Carlo), and 0.74 (5-Fold) further demonstrate the model's reliability in handling the minority class (true matches). The chance line in the ROC plot represents random classification, included for reference.

While metrics like precision, recall, and F1 score depend on threshold tuning, PR-AUC and AUC offer threshold-independent evaluations of the model's overall performance. PR-AUC, in particular, is well-suited for imbalanced datasets as it focuses on the minority class; true matches; where performance matters most. The model's PR-AUC of 0.75 reflects a reasonable ability to balance precision and recall under the challenging conditions of an imbalanced dataset. In contrast, AUC, which reflects the model's overall discriminative power, is less sensitive to class imbalance but reinforces confidence in the model's ability to distinguish between matches and non-matches across varying thresholds.

The F1 and F2 scores, ~ 0.68 and 0.69 , respectively, indicate reasonable performance in forensic contexts. These metrics balance precision and recall, with the F2 score

emphasizing recall to ensure true matches are not overlooked. This trade-off is particularly valuable in forensic investigations, where both false positives and false negatives carry significant consequences.

Collectively, these metrics highlighted the model's strengths in providing a robust framework for bite mark analysis. Accuracy, specificity, and AUC underscored its strong general reliability and discriminative ability, while PR-AUC, F1, and F2 scores emphasized its effectiveness in processing imbalanced data. Despite moderate precision and recall, the model offers forensic investigators a reliable tool to identify suspects effectively, minimizing the risks of wrongful identification or missed matches. With potential enhancements, such as a larger dataset, these metrics could further improve, bolstering the model's reliability and applicability in forensic scenarios.

4 Discussion

4.1 Study discussion and context

The primary objective of this study was to explore the relationship between teeth dentition and bite mark analysis, with a focus on developing standardized, data-driven methodologies to improve the accuracy and reliability of forensic bite mark comparisons. Research in this area is crucial given the variability and subjectivity that have historically plagued bite mark analysis. Christoloukas et al. (15) emphasized this need in their comprehensive review of all published articles on bite mark analysis from 2012 to 2023. They concluded that while bite mark analysis has the potential to be a reliable complementary tool in forensic investigations, its effectiveness hinges on the adoption of scientifically validated methodologies, standardized protocols for data collection and processing, and clear evaluation criteria. This underscores the importance of establishing rigorous and universally accepted methods—the focus of the present research. By addressing these gaps, this study aims to contribute to the development of a more objective and scientifically grounded framework for bite mark analysis.

4.2 Composite scoring model and Principal Component Analysis (PCA)

The findings of this study aligned with the observations made by Sweet and Pretty (7) regarding the evidentiary value of bitemarks. They noted that in cases where distinctive individual characteristics of teeth are recorded, bitemarks can be a reliable source of forensic evidence. The model aligned with observations by Sweet and Pretty (7) on the evidentiary value of bite marks, particularly when distinctive individual characteristics of teeth were recorded. This study built on these insights by introducing a composite scoring model using

Principal Component Analysis (PCA), which aimed to reduce subjective variability and enhance the accuracy of matching dental impressions to bitemark evidence.

The results of this study emphasized the value of using Principal Component Analysis (PCA) as an advanced tool in forensic odontology, particularly for bite mark analysis. By introducing a standardized approach for dental impression and bite evidence comparison, this research addressed the challenge of subjectivity in bite mark identification, offering a more consistent and data-driven methodology. Traditionally, bite mark analysis has been criticized for its susceptibility to examiner bias and inconsistent practices, which have led to discrepancies in legal cases and, at times, wrongful convictions. This new approach provided an evidence-based solution by transforming bite mark identification into a quantitative process that relied on measurable features, such as surface area, inter-molar and inter-canine distances, and rotational angles of teeth.

The performance metrics observed in this study demonstrated the effectiveness of the composite scoring system in bite mark analysis. Precision values of 68% in the 5-fold cross-validation and 69% in Monte Carlo cross-validation indicated that the model generally minimized false positives, which is crucial in forensic science to avoid misidentifying individuals. Similarly, recall rates of 69% in the 5-fold cross-validation and 70% in Monte Carlo cross-validation reflected the model's ability to correctly identify true matches, reducing the likelihood of missing critical evidence. The F1 scores, ranging from 68% in 5-fold cross-validation to 69% in Monte Carlo cross-validation, further underscored a balanced performance between

precision and recall, helping manage both false positives and false negatives effectively.

The high accuracy values of 91% in 5-fold cross-validation and 92% in Monte Carlo cross-validation underscored the model's strong predictive capability for bite mark analysis, reflecting its reliability in correctly identifying matches across varied subsets. While accuracy provides an acceptable general measure of performance; precision, recall, and F1 scores further underscored the model's balanced approach to both identifying true matches and minimizing false positives—two critical factors in forensic applications.

Specificity values of 95% across both 5-fold and Monte Carlo cross-validation confirmed that the model was adequate at correctly identifying non-matches, thereby reducing the risk of false positives. The Area Under the Curve (AUC) value of 0.92 in both validation methods reflected a strong discriminative ability, showcasing the model's effectiveness in distinguishing between true matches and non-matches across various thresholds. These results underscored the model's robustness in forensic applications, providing a reliable tool for identifying potential matches in bite mark analysis.

4.3 Comparison with previous studies

This study's approach and performance metrics aligned with, and in some cases exceeded, those reported in prior research on forensic bite mark analysis. A key innovation introduced in this study was the use of PCA-based standardization to develop a composite scoring model, an approach not previously applied in bite mark analysis. By anchoring the scoring system in objective, quantifiable dental features, this method aimed to reduce examiner subjectivity, providing a more consistent framework than

earlier models that relied on subjective assessments.

For instance, Pretty and Sweet (16) documented high error rates due to subjective biases, with false-positive rates reaching 45.5% and false-negative rates up to 71.4%. In contrast, our PCA-driven composite scoring method addressed such biases by objectively weighting features, as reflected in its strong performance metrics.

The model's AUC value of 0.92 demonstrated significant discriminatory power, surpassing or matching other models in the field. For example, Martin-De-Las-Heras et al. (17) used a combination of 3D-scanned dental casts and logistic regression, achieving ROC values between 0.8 and 0.9. The comparable AUC achieved by our PCA-based approach suggested a similarly robust ability to distinguish true matches from non-matches. Additionally, Fournier et al. (18) introduced intraoral scanners and mesh comparison software for bite mark analysis; however their approach still relied on manual evaluation of color gradients for final assessment, making it susceptible to subjectivity. In contrast, our model leveraged PCA standardization to provide a more robust, quantitative scoring system, which aimed to enhance consistency across different forensic cases.

Molina et al. (19) also utilized 3D scans of dental casts, focusing on key dental measurements such as inter-canine distances and angular positions. However, their model, which relied on Euclidean distances, achieved a maximum ROC value of 0.73, indicating lower discriminatory power compared to the approach in this study, which attained AUC values of 0.92. This contrast suggested that integrating composite scores derived from multiple

features, as done in this study, may provide a more robust and accurate assessment, enhancing the model's ability to reliably distinguish true matches from non-matches in forensic bite mark analysis.

Finally, Tarvadi et al. (20) attempted to improve bite mark analysis by collecting 14 individual measurements for each bite mark and summing them to find a representative score. However, this approach, despite its detailed data collection, resulted in an accuracy of only 28%; significantly lower than the results achieved in this study and other comparable research. This low accuracy underscores the limitations of relying solely on simple aggregated measurements and highlights the need for more robust statistical methodologies and composite scoring systems to enhance the reliability of bite mark analysis.

4.4 Variability in bite mark analysis and model limitations

Despite efforts in standardizing measurements, inherent variability continued to impact the accuracy of bite mark analysis. Sweet and Pretty (16) highlighted significant concerns regarding examiner bias and measurement reliability in bite mark comparisons. Their study, which evaluated digital overlays for comparing known and questioned bite marks, found elevated levels of examiner error. In particular, they found intra-examiner agreement (where an examiner's judgments remained consistent with their previous assessments) as low as 65%, indicating notable variability in individual performance. Additionally, they reported an average false-positive rate of 15.9%, with instances reaching as high as 45.5%, and a false-negative rate averaging 25.0%, with some cases reaching up to 71.4%. These findings underscore the susceptibility of traditional bite mark analysis to examiner bias and

inconsistency, factors that can heavily influence the reliability of forensic conclusions.

To address these challenges, this study implemented a methodology involving five independent measurements per sample by different examiners, with each measurement repeated across five trials to reduce measurement error. This multi-examiner approach aimed to mitigate individual biases by averaging the measurements across both examiners and trials. Additionally, any measurements identified as outliers were excluded from the calculation, further enhancing the accuracy and reliability of the results. This combined strategy of multi-examiner input, repeated trials, and outlier exclusion minimized personal biases and measurement error, contributing to a more standardized and objective framework for bite mark analysis.

Moreover, external factors like tissue inflammation, decomposition, and positional shifts are known to distort bite marks, adding complexity to the identification process. For instance, McDowell (21) highlighted how these variables influenced bite mark appearance and complicated comparisons. Additionally, positional changes of the victim and movements from both the biter and the bitten can introduce unquantifiable variability, further challenging the accurate identification of a specific biter.

Given these known sources of error, future studies are essential to assess their impact on bite mark distortion and to refine the composite scoring model introduced in this study. A deeper understanding of these factors and their effects on forensic analyses will be critical for enhancing the reliability and scientific validity of bite mark evidence.

4.5 Limitations and perspectives

While this study established a strong foundation, addressing certain limitations and exploring additional methodological improvements could further enhance the model's applicability and robustness in forensic contexts. Given the moderate sample size, future research should focus on expanding the dataset to improve statistical robustness and generalizability. A larger dataset would support a more representative sampling across demographic groups, enabling the model to consider variables such as gender, ethnicity, and age when recalculating composite scores and establishing thresholds. With a limited sample size, capturing a broad representation of the population is challenging, constraining the model's ability to account for demographic diversity. Dental characteristics have been shown to vary significantly across demographic categories, potentially influencing the bite mark matching process. Future iterations of the model should evaluate whether different thresholds and scoring systems are necessary for different demographic groups to maintain accuracy and reliability. Tailoring the model to incorporate these factors could enhance precision and support equitable forensic practices.

This study's use of high-resolution cameras and standard software tools like Adobe Photoshop proved effective, though factors such as lighting conditions and image resolution may affect measurement precision. Future studies should prioritize standardized lighting and equipment calibration to minimize potential errors. Exploring advanced imaging techniques and specialized equipment could further improve data quality, resulting in more accurate forensic measurements.

To increase the objectivity of bite mark analysis, future research should incorporate blinded testing protocols, where examiners are unaware of identifying details about samples. Such protocols would reduce examiner bias, providing an impartial basis for comparisons and strengthening the model's reliability.

Additional promising directions for advancing bite mark analysis include integrating machine learning algorithms for automated pattern recognition, which could streamline the analysis process. Ensemble methods, such as consensus or voting based scoring across different scoring methods, may boost the robustness of the scoring process. Finally, exploring non-linear PCA and regression-based weighting techniques could capture complex relationships within the data, optimizing feature representation and composite scoring accuracy.

Addressing these limitations and following these future directions will support a more objective, reliable, and scientifically rigorous framework in forensic bite mark analysis, helping to standardize practices and improve confidence in forensic evidence.

5 Conclusion

By developing a composite scoring system based on PCA, this study provided a replicable framework that was both objective and grounded in measurable dental features. The performance metrics achieved with this method highlighted its potential for adoption as a standard practice in forensic odontology. Specifically, the model demonstrated a high AUC of 0.92 and a specificity of 0.95, indicating its strong discriminatory power and reliability in accurately identifying non-matches; critical in minimizing false positives in forensic contexts. The PR-AUC of 0.75

further highlighted the model's ability to balance precision and recall, making it particularly effective for addressing the challenges of imbalanced forensic datasets. Precision and recall values of ~ 0.67 - 0.69 indicated a balanced approach to identify true matches while minimizing false positives. However, as with many studies in this domain, the moderate precision and recall values, along with the limited sample size, underscore the importance of further research to expand the dataset and enhance model performance. Larger and more diverse datasets would be instrumental in improving the model's robustness and ensuring its applicability across varied populations and complex bite mark cases.

The introduction of an acceptance threshold, derived through a grid search of performance metrics, provided a heuristic yet effective filter for matching bite marks to dental molds. The grid-searched threshold of 12.65, confirmed through cross-validation, balanced precision and recall, making it suitable for forensic applications where specificity is critical. The decision to standardize data before applying PCA also played a crucial role in maintaining balanced input contributions, preventing features measured on larger scales from disproportionately influencing the analysis, ensuring consistent and reliable performance.

In conclusion, this composite scoring system lays a promising foundation for a more objective, scientifically rigorous approach to bite mark analysis. As forensic science advances, the adoption of standardized, data-driven methods will play an increasingly important role in upholding the reliability and credibility of forensic evidence in legal proceedings. Continued research and refinement of this framework can further increase its value, offering forensic professionals a dependable tool for accurate, unbiased, and equitable bite mark analysis.

Future research should focus on expanding the dataset, testing the composite scoring approach across a variety of bite mark materials, and refining the threshold criteria to achieve even greater accuracy and reliability.

6 Acknowledgements

I would like to extend my sincere gratitude to Ms. Kate Allender from Tesla STEM High School for her unwavering support and encouragement throughout this project. I would like to extend my gratitude to the volunteers who participated in this study and my peers who assisted in taking the dental casts. I am also thankful to the American Society of Forensic Odontology, the American Academy of Forensic Sciences, and the Association Forensic Odontology for Human Rights for their invaluable guidance and support throughout this research.

References

1. American Board of Forensic Odontology (ABFO). (2017). Body Identification Information & Guidelines. <https://abfo.org/wp-content/uploads/2012/08/ABFO-Body-ID-Information-Guidelines-Feb-2017.pdf>
2. Avon, S. L. (2004). Forensic Odontology: the roles and responsibilities of the dentist. In *Journal of the Canadian Dental Association* 70(7), 453–454. <https://www.cda-adc.ca/jcda/vol-70/issue-7/453.pdf>
3. Jayakrishnan, J. M., Reddy, J., Vinod Kumar, R. B. (2021). Role of forensic odontology and anthropology in the identification of human remains. *Journal of Oral and Maxillofacial Pathology (JOMFP)*, 25(3), 543–547. https://doi.org/10.4103/jomfp.jomfp_81_21
4. Gupta, K. (2017). Advantages and disadvantages of forensic odontology. *International Education and Research Journal (IERJ)*, 3(5) <https://ierj.in/journal/index.php/ierj/article/view/1032>
5. Pittayapat, P., Jacobs, R., De Valck, E., Vandermeulen, D., Willems, G. (2012). Forensic odontology in the disaster victim identification process. *The Journal of forensic odontology*, 30(1), 1–12. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5734849/>
6. Mohammed, F., Fairizekhan, A. T., Bhat, S., Menezes, R. G. (2023, August 14). *Forensic Odontology*. StatPearls - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK540984/>
7. Pretty, I., Sweet, D. (2001a). The scientific basis for human bitemark analyses – a critical review. *Science & Justice*, 41(2), 85–92. [https://doi.org/10.1016/s1355-0306\(01\)71859-x](https://doi.org/10.1016/s1355-0306(01)71859-x)
8. Rawson, R. D., Ommen, R. K., Kinard, G., Johnson, J., Yfantis, A. (1984). Statistical evidence for the individuality of the human dentition. *Journal of Forensic Sciences*, 29(1), 245–253. <https://doi.org/10.1111/j.1556-4029.2010.01531.x>
9. Bush, M. A., Bush, P. J., Sheets, H. D. (2010). Statistical evidence for the similarity of the human dentition. *Journal of Forensic Sciences*, 56(1), 118–123. <https://doi.org/10.1111/j.1556-4029.2010.01531.x>
10. Franco, A., Willems, G., Souza, P., Tanaka, O., Coucke, W., Thevissen, P. (2017). Three-dimensional analysis of the uniqueness of the anterior dentition in orthodontically treated patients and twins. *Forensic Science International*, 273, 80–87. <https://doi.org/10.1016/j.forsciint.2017.02.010>

11. Sauerwein, K., Butler, J. M., Reczek, K. K., Reed, C., Division 602 - Special Programs Office, Division 601 – Standards Coordination Office, National Institute of Standards and Technology, & U.S. Department of Commerce. (2023). Bitemark Analysis: A NIST Scientific Foundation review. In *NIST Interagency Report*. <https://doi.org/10.6028/NIST.IR.8352>
12. Barsley, R. E., Bernstein, M. L., Brumit, P. C., Dorion, R. B., Golden, G. S., Lewis, J. M., et al. (2018). Epidermis and enamel. *American Journal of Forensic Medicine & Pathology*, 39(2), 87–97. <https://doi.org/10.1097/paf.0000000000000392>
13. Saks, M. J., Albright, T., Bohan, T. L., Bierer, B. E., Bowers, C. M., Bush, M. A., et al. (2016). Forensic bitemark identification: weak foundations, exaggerated claims. *Journal of Law and the Biosciences*, 3(3), 538–575. <https://doi.org/10.1093/jlb/lsw045>
14. Rea, A., Rea, W. (2016, October 12). *How Many Components should be Retained from a Multivariate Time Series PCA?* arXiv.org. <https://doi.org/10.48550/arXiv.1610.03588>
15. Christoloukas, N., Mitsea, A., Rontogianni, A., Papadakis, E., Angelopoulos, C. (2024). Evaluation of Bitemark Analysis's potential application in Forensic Identification: A Systematic review. *Diagnostics*, 14(11), 1180. <https://doi.org/10.3390/diagnostics14111180>
16. Pretty, I. A., Sweet, D. (2001b). Digital Bite Mark Overlays—An Analysis of Effectiveness. *Journal of Forensic Sciences*, 46(6), 15160J. <https://doi.org/10.1520/jfs15160j>
17. Martin-De-Las-Heras, S., Tafur, D., Bravo, M. (2013). A quantitative method for comparing human dentition with tooth marks using three-dimensional technology and geometric morphometric analysis. *Acta Odontologica Scandinavica*, 72(5), 331–336. <https://doi.org/10.3109/00016357.2013.826383>
18. Fournier, G., Savall, F., Nasr, K., Telmon, N., Maret, D. (2019). Three-dimensional analysis of bitemarks using an intraoral scanner. *Forensic Science International*, 301, 1–5. <https://doi.org/10.1016/j.forsciint.2019.05.006>
19. Molina, A., Ramos, B., Torres, J. C., Martin-De-Las-Heras, S., Perez, M. B. (2020). Dental parameter quantification with semi-automatized computational technology for the analysis of human bitemarks. *Australian Journal of Forensic Sciences*, 54(2), 247–257. <https://doi.org/10.1080/00450618.2020.1793385>
20. Tarvadi, P. V., Manipady, S., Shetty, M. (2016). Bite marks analysis using metric method. *Indian Journal of Forensic Medicine & Toxicology*, 10(2), 277. <https://doi.org/10.5958/0973-9130.2016.00111.0>
21. McDowell, J. D. (2023). Is it likely that bite mark analysis will in the future meet the criteria necessary for use in the justice system? *Journal of the California Dental Association*, 51(1). <https://doi.org/10.1080/19424396.2023.2210328>