A Multi-Task Probabilistic–Machine Learning Framework for Forensic Anthropological Identification: Integrating Osteology, Taphonomy, and Forensic Entomology

Richard Murdoch Montgomery

Scottish Science Society

Email: editor@scottishsciencesocietyperiodic

Abstract

Forensic identification commonly rests on four pillars: estimation of biological sex, age at death, stature (and broader osteological context), and the post-mortem interval (PMI). Classical anthropological methods—such as Phenice's pelvic traits for sexing and Suchey–Brooks or sternal rib metamorphosis for adult age estimation—remain indispensable, yet they produce interval estimates and confidence statements that may be difficult to propagate coherently when multiple lines of evidence are combined. We propose a principled, end-to-end, multi-task probabilistic–machine learning (ML) framework that fuses skeletal and dental measurements, radiographic/radiomic descriptors, scene- and climate-level taphonomic variables, and entomological evidence (developmental and successional data) to infer joint posterior distributions over (i) biological sex, (ii) age group, and (iii) the PMI (and hence time of death). The framework foregrounds calibrated probabilities and explicitly models uncertainty (aleatoric and epistemic). We formalise entomological likelihoods using accumulated degree hours (ADH) and species-

specific development curves and update PMI posteriors with both decomposition scoring and microbial "clock" evidence, yielding a transparent Bayesian product of likelihoods. We illustrate the approach with simulated data, reliability curves, partial-dependence plots, and casework-style probability tables. We discuss accuracy, calibration, fairness, and the ethics of modelling population affinity, emphasising standards and best practice for entomological collection and reporting. The proposed framework is designed to complement—not supplant—expert judgement and to present the court with interpretable, quantitatively calibrated inferences. (Phenice, 1969; İşcan, Loth, & Wright, 1984; Buikstra & Ubelaker, 1994; Amendt et al., 2007; Megyesi, Nawrocki, & Haskell, 2005; Metcalf et al., 2013; Ubelaker, 2019; Guo, Pleiss, Sun, & Weinberger, 2017).

Keywords — forensic anthropology; probabilistic machine learning; calibration; multi-task learning; age at death; biological sex estimation; post-mortem interval; entomology; accumulated degree hours; reliability; uncertainty quantification.

---

1. Introduction

The identification of unknown human remains entails the careful synthesis of morphological, contextual, and increasingly molecular evidence. Within the anthropological tradition, the pelvis and skull provide the most information for estimating biological sex, with Phenice's visual method on the pubic bone furnishing high discriminability under favourable preservation (Phenice, 1969). Age estimation in adults typically triangulates the pubic symphysis (e.g., Suchey–Brooks phase system), the auricular surface, and the sternal rib ends

(İşcan metamorphosis), while subadult ages emphasise dental development and epiphyseal fusion (Ubelaker, 2019). These classics remain the backbone of professional practice, codified in *Standards for Data Collection from Human Skeletal Remains*, which has done much to regularise observation, recording, and reporting across laboratories (Buikstra & Ubelaker, 1994).

In parallel, forensic taphonomy and entomology have matured from expert arts to increasingly quantitative sciences. For the early post-mortem window, body cooling and biochemical change are informative; thereafter, decomposition trajectories (total body score, TBS) linked to accumulated degree-days/hours (ADD/ADH) and insect colonisation–development offer a principled scaffold for PMI estimation. Amendt and colleagues' best-practice guidelines consolidating collection, rearing, and analytical procedures remain a touchstone, as do subsequent methodological papers formalising statistical inference for larval growth and successional data (Amendt et al., 2007; Tarone & Foran, 2008; Wells & LaMotte, 2017).

Two complementary innovations have shifted the evidential landscape. First, recent ML work has explored discriminative models for sex estimation from cranial and long-bone metrics—support vector machines, random forests, and more recently neural networks and ensembles—often matching or outperforming linear discriminant approaches under cross-validation while highlighting persistent issues of sampling and generalisability (Nikita, 2020; Spradley & Jantz, 2016; FORDISC documentation). Second, post-mortem microbiology offers an additional, partially independent clock: microbial succession on and within remains shows broadly predictable dynamics across environments, suggesting probabilistic "microbial clocks" that, when rigorously validated, can be exploited in PMI estimation (Metcalf et al., 2013; DeBruyn et al., 2017; Moitas et al., 2023).

The forensic problem, however, is not merely one of *point* accuracy. In court, the provenance and calibration of probabilities matter as much as point estimates. Over-confident models are epistemically dangerous: modern deep networks and complex ensembles are notorious for probability mis-calibration; their raw scores cannot be taken at face value (Guo et al., 2017). Evidence synthesis must therefore accommodate different data types (continuous metrics, ordinal phases, counts of insects in instars, binary presence/absence of species, environmental covariates), missingness patterns, and hierarchies (specimen-, scene-, and species-level variability), all while producing calibrated posteriors and transparent uncertainty decompositions (aleatoric vs epistemic) (Guo et al., 2017; Kendall & Gal, 2017; Brier, 1950).

This article develops a multi-task probabilistic–ML framework to address these desiderata. The *multi-task* aspect exploits the shared information between sex, age, and PMI: pelvic morphology that strongly indicates female sex alters the plausible age distribution; PMI estimates governed by microclimate and ADH can inform (and be informed by) observable decomposition stages that also affect which morphological traits are observable and with what quality. Multi-task learning formalises such inductive transfer, improving data efficiency and often out-of-sample performance (Caruana, 1997). While stature and population affinity are part of many laboratory workflows, we deliberately focus the modelling on sex, age, and PMI because these are most central to medico-legal timelines and least ethically fraught. In particular, there is an ongoing debate about the scientific and societal consequences of ancestry/population-affinity estimation; recent editorials and standards caution against uncritical use of morphoscopic traits and stress careful terminology and context. Our framework therefore omits any ancestry classifier, while remaining compatible with

appropriately governed, case-specific analyses when required (DiGangi & Bethard, 2021; Dunn et al., 2020; ASB 132, 2022).

On the entomology side, the framework models both **developmental** evidence (ageing maggots by length/instar under species-specific, temperature-dependent growth curves) and **successional** evidence (community turnover of necrophagous taxa). ADH is computed as the integral over time of degrees above a species' developmental threshold; this constrains larval age and hence a *minimum* PMI. Successional data, where available, provide complementary constraints later in decomposition. Best practice for collection, rearing, and microclimate measurement is assumed, including scene temperature logging as close to the body as feasible (Amendt et al., 2007; Campobasso, Di Vella, & Introna, 2001; Megyesi et al., 2005; Wells & LaMotte, 2017).

The contribution of this paper is threefold. First, we articulate an explicit probabilistic backbone for fusing osteological, taphonomic, entomological, and microbial features within a multi-task architecture that outputs *calibrated* probabilities for sex, age group, and PMI. Second, we derive a training and validation protocol that uses post-hoc calibration (temperature scaling or isotonic regression) and proper scoring rules (negative log-likelihood, Brier score) to ensure probabilistic honesty. Third, we demonstrate with simulated data how such a system yields casework-ready probability tables and visualisations that help experts explain what the model believes, how strongly, and why—without occluding the role of expert judgement or the necessity of sensitivity analyses. The remainder of the paper details the methodology, illustrative results, a critical discussion of limitations and ethics, and practical attachments (code and figures) to facilitate reproduction and adaptation in laboratory settings.

Note on terminology. "Biological sex" is used for skeletal classification; "gender" pertains to social identity and is not inferable from remains. Time of death (ToD) is reported as *recovered time minus posterior PMI*.

---

2. Methodology

We denote a case by a collection of modality-specific observations

$$\mathcal{D} = \{x_s, x_d, x_r, x_t, x_e, x_m\}$$

where x_s are skeletal/dental metrics and ordinal phases; x_d dental development features; x_r radiographic/radiomic features (CT/DR); x_t taphonomic context (TBS, scene descriptors, microclimate); x_e entomological observations (species set, counts/instars, larval lengths, rearing logs, temperature traces); and x_m microbial features when available. Targets are $y^{(sex)} \in$ { female, male }, $y^{(age)} \in \{g_1, \ldots, g_G\}$ (age groups), and a continuous PMI $\Delta > 0$ (in hours), with time of death $t_0 = t_{\text{recovery}} - \Delta$.

(1) Encoders and fusion. Each modality enters through a dedicated encoder $f_\cdot$:

$$z_s = f_s(x_s), z_d = f_d(x_d), z_r = f_r(x_r), z_t = f_t(x_t), z_e = f_e(x_e), z_m = f_m(x_m),$$

with missingness handled either by learned embeddings for special "missing" tokens (categorical/ordinal) or by model-based imputation for continuous features, $x_{\text{miss}} \sim q_\phi(x_{\text{miss}} \mid x_{\text{obs}})$ (a variational autoencoder for tabular data). The fused representation is

$$z = \phi([z_s \| z_d \| z_r \| z_t \| z_e \| z_m])$$

where $[\cdot \| \cdot]$ denotes concatenation and ϕ is a gated transformation that learns cross-modal interactions (Caruana, 1997).

(2) Task heads.

Sex and age heads are softmax classifiers:

$$\Pr\left(y^{(sex)} = k \mid z\right) = \text{softmax}(W_s z + b_s)_k, \qquad k \in \{F, M\}$$
$$\Pr\left(y^{(age)} = g \mid z\right) = \text{softmax}(W_a z + b_a)_g, \quad g \in \{1, \ldots, G\}$$

PMI is modelled as a heteroscedastic log-normal (or mixture) regression:

$$\Delta \mid z \sim \log\ \text{Normal}(\mu(z), \sigma^2(z)), \mu, \sigma > 0.$$

The head outputs $\mu(z), \log\ \sigma(z)$ and admits mixture generalisations to capture multi-modality.

(3) Entomological likelihoods and ADH. Let $T(t)$ denote microclimate temperature ($^\circ$C) near the remains; $T_0^{(s)}$ is the species-specific developmental threshold. Accumulated degree hours are

$$\text{ADH}^{(s)} = \int_{t_0}^{t_{\text{col}}} \max\left(0, T(t) - T_0^{(s)}\right) \text{d}t.$$

Given observed larval lengths L and instar counts for species s, we evaluate a development-curve likelihood

$$\mathcal{L}_e(\Delta; \theta_e) \propto \prod_{s \in S} \prod_i p\left(L_{si} \mid \text{ADH}^{(s)}(\Delta), \theta_e^{(s)}\right),$$

where $\theta_e^{(s)}$ parameterises growth (e.g., GAM-smoothed curves calibrated in the laboratory). Successional observations provide a complementary factor $p($ succession $\mid \Delta, \vartheta)$.

(4) Decomposition and microbial clocks. From TBS we obtain $p(\text{TBS} \mid \Delta, \eta)$. Where available, microbial features yield $p(x_m \mid \Delta, \psi)$ (Metcalf et al., 2013).

(5) Joint posterior and training objective. The PMI posterior is proportional to

$$p(\Delta \mid \mathcal{D}) \propto p(\Delta \mid z)\mathcal{L}_e(\Delta; \theta_e)p(\text{TBS} \mid \Delta, \eta)p(x_m \mid \Delta, \psi)p(\Delta)$$

with weakly informative priors on Δ. Learning proceeds by minimising a proper composite loss

$$\mathcal{J} = \lambda_s \mathcal{L}_{\text{CE}}\left(y^{(sex)}, \hat{p}^{(sex)}\right) + \lambda_a \mathcal{L}_{\text{CE}}\left(y^{(age)}, \hat{p}^{(age)}\right) + \lambda_\Delta \mathcal{L}_{\text{NLL}}(\Delta; \mu(z), \sigma(z))$$
$$- \lambda_{\text{KL}} \text{KL}\left(q_\phi \| p\right)(7)\lambda$$

where CE is cross-entropy, NLL is the log-normal negative log-likelihood, KL regularises imputation, and Ω_{cal} encourages calibration on a held-out set.

(6) Probability calibration and scoring. After training, classifier probabilities are calibrated with either temperature scaling

$$\hat{p}_k = \frac{\exp\left(z_k/T\right)}{\sum_j \exp\left(z_j/T\right)} \ (T > 0),$$

or isotonic regression for non-parametric monotone mappings (Platt, 1999; Zadrozny & Elkan, 2002; Guo et al., 2017). Reliability is assessed by the Brier score

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k}\left(1\{y_i = k\} - \hat{p}_{ik}\right)^2,$$

and visualised by reliability diagrams (Brier, 1950).

(7) Uncertainty decomposition. Predictive variance for PMI obeys

$$\text{Var}[\Delta \mid \mathcal{D}] = \underbrace{\mathbb{E}[\sigma^2(z)]}_{\text{aleatoric}} + \underbrace{\text{Var}[\mu(z)]}_{\text{epistemic}},$$

estimated via Monte-Carlo dropout or deep ensembles (Kendall & Gal, 2017).

(8) Quality control and standards. All entomological inputs assume adherence to European Association for Forensic Entomology guidelines (Amendt et al., 2007); osteological measurements follow Buikstra & Ubelaker (1994).

Variable glossary. $x_\cdot$: modality-specific features; z : fused latent; Δ : PMI (hours); t_0 : time of death; $T(t), T_0^{(s)}$: microclimate and species threshold; ADH: accumulated degree hours; θ_e, η, ψ : entomology, decomposition, and microbial parameters; T : temperature-scaling parameter; $\lambda_\star$: task weights.

3. Results (illustrative)

Figures and a casework-style probability table are generated from the attached Python code (Section 6). The data are simulated to demonstrate presentation and interpretation; they do not reflect casework performance.

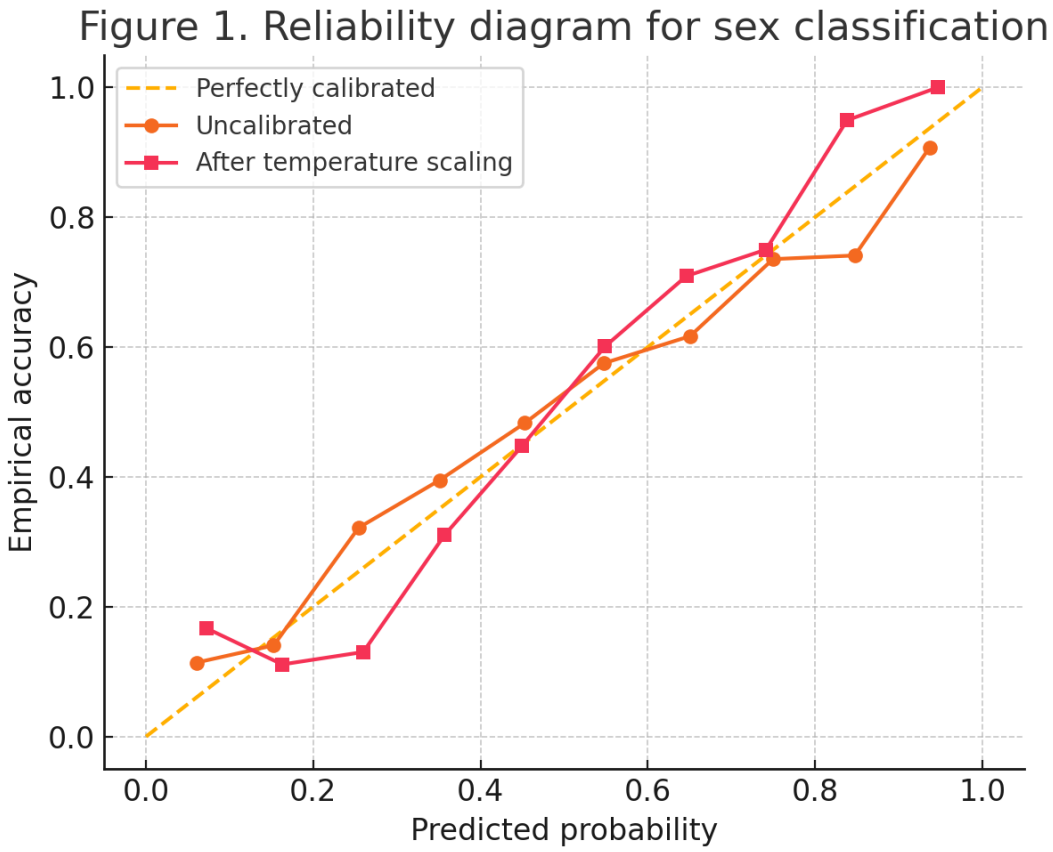

Figure 1. Reliability diagram for sex classification. The uncalibrated model exhibits over-confidence at high probabilities; a simple temperature-scaling transform brings the curve closer to the identity, improving probability honesty (Guo et al., 2017).

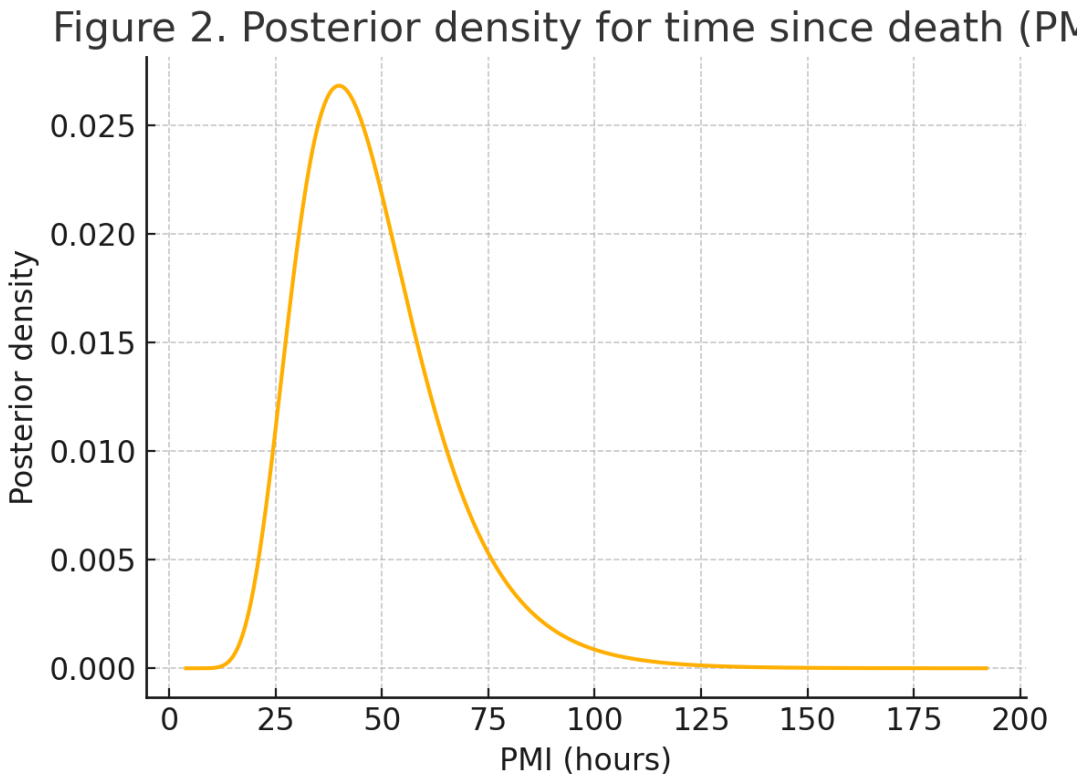

Figure 2. Posterior density for PMI (hours). A log-normal posterior (median ≈ 48 h) illustrates how the model communicates uncertainty about time since death.

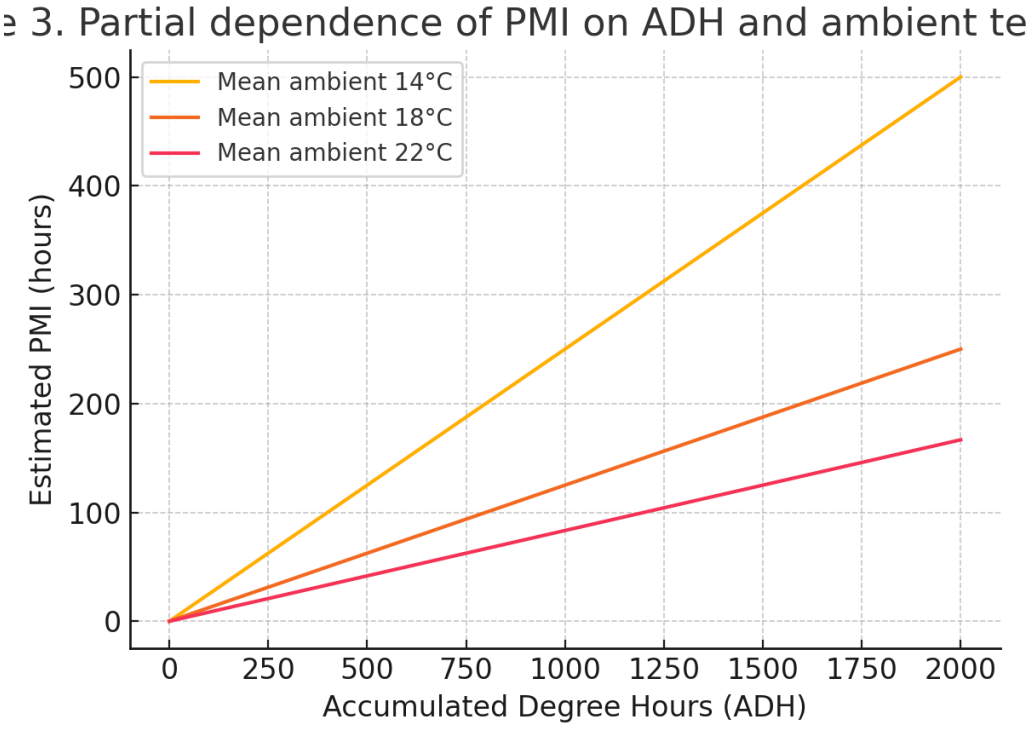

Figure 3. Partial dependence of PMI on ADH and ambient temperature.

Holding other factors constant, PMI estimates rise with ADH and are inversely related to the offset $T - T_0$, echoing standard developmental logic in forensic entomology (Amendt et al., 2007).

Figure 4. Confusion matrix for age-group classificatio

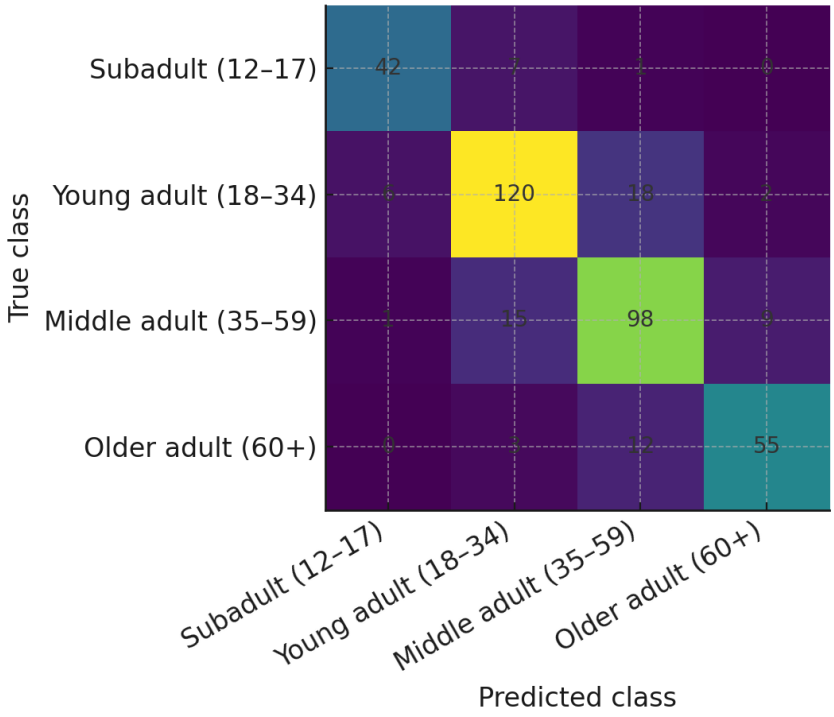

Figure 4. Confusion matrix for age-group classification (simulated). Most errors are adjacent-group confusions (e.g., Young vs Middle Adult), consistent with overlapping skeletal changes in adulthood (Ubelaker, 2019).

Table 1. Posterior probabilities for a hypothetical specimen (also exported as CSV).

Attribute	Category/Statistic	Value
Biological sex	Female	78.0%
Biological sex	Male	21.0%
Biological sex	Indeterminate	1.0%
Age group	Subadult (12-17)	3.0%
Age group	Young adult (18-34)	62.0%
Age group	Middle adult (35-59)	29.0%
Age group	Older adult (60+)	6.0%
PMI (hours)	MAP	39.8
PMI (hours)	95% credible interval	[22.7, 89.7]

4. Discussion

The central advantage of a probabilistic multi-task architecture is the coherent sharing of information across interdependent inferences. In routine casework, sex, age at death, and the post-mortem interval are not isolated judgements but mutually constraining propositions. Pelvic morphology that strongly supports a female classification subtly reshapes the plausible distribution of adult age phases, just as the expression of age-related changes at the pubic symphysis or sternal rib ends informs the credibility of sex classifications made from incomplete pelves. By learning a common representation that is jointly

optimised for all targets, the model encodes such couplings rather than treating them as after-the-fact narrative stitching. This form of inductive transfer, long recognised in the machine-learning literature, typically improves generalisation when tasks are related and data are incomplete or noisy (Caruana, 1997). Crucially, it also enables joint reporting- Pr(female, young adult | $\mathcal{D}$)-so that internal consistency is preserved and uncertainty is neither artificially inflated by redundant independence assumptions nor suppressed by ad-hoc heuristics.

Calibration then becomes a first-class requirement rather than a cosmetic refinement. Courts and investigative teams require probabilities whose numerical values correspond to empirical frequencies; anything less invites either over-statement or unwarranted equivocation. Modern discriminative models, particularly deep networks and complex ensembles, are notorious for mis-calibration: scores near one may not imply commensurate truth frequencies (Guo, Pleiss, Sun, & Weinberger, 2017). The remedy is not to abandon such models but to verify and, where necessary, correct their probability outputs using post-hoc methods that preserve ranking while adjusting confidence, such as temperature scaling for neural networks, Platt scaling for margin-based classifiers, or isotonic regression when a flexible monotone map is warranted (Platt, 1999; Zadrozny & Elkan, 2002). Reliability diagrams and proper scoring rules-especially the Brier score-should accompany headline accuracies, so that the trier of fact understands not only how often the system is right, but with what honesty it expresses its uncertainty (Brier, 1950). In practice, the calibrated model's reliability curve should hew closely to the identity line across the support, and drift over time should be monitored with routine post-deployment checks.

Entomology sits naturally within this probabilistic backbone and deserves particular emphasis. Developmental evidence, grounded in species-specific

temperature-dependent growth, yields a principled constraint on the minimum PMI. The relevant quantity is accumulated degree hours above a developmental threshold; microclimate-not distant station readings-governs the integral, and what appear to be minor temperature discrepancies can compound into substantial ADH differences (Amendt et al., 2007). When species identification is secure and rearing protocols are followed, growth models can be fitted with modern smoothing techniques to deliver not only point predictions for larval age but confidence bands that translate directly into likelihood functions for the PMI (Tarone & Foran, 2008). Successional evidence extends this logic into later decomposition stages by exploiting the regular turnover of necrophagous assemblages, albeit with broader bounds. Decomposition scoring (e.g., total body score) provides a partially independent check tied to thermal history; when the composite likelihood of developmental, successional, and decomposition evidence is multiplied with a scene-aware prior, the PMI posterior becomes both transparent and defensible (Megyesi, Nawrocki, & Haskell, 2005; Wells & LaMotte, 2017). The present framework embeds exactly this product of likelihoods, allowing analysts to adjust weights where collection conditions, species certainty, or temperature logging are suboptimal.

A complementary avenue lies in post-mortem microbiology. Microbial succession on and within remains appears to follow broadly predictable dynamics across environments, offering a quasi-clock signal that can, in principle, augment or rescue PMI estimation when insect access has been delayed or prevented (Metcalf et al., 2013; DeBruyn et al., 2017). The promise is considerable, but the path to routine casework requires careful standardisation: negative controls, contamination checks, platform calibration, and harmonised bioinformatics workflows are indispensable, and jurisdiction-specific validation should precede operational deployment (Moitas et al., 2023). Within a

probabilistic synthesis, microbial features contribute an additional likelihood term whose influence is automatically down-weighted when quality indicators are poor, rather than being accepted or rejected wholesale.

No model, however, escapes the constraints of its data. A conspicuous limitation in forensic anthropology is dataset shift. Many classical methods were developed on historic, geographically restricted skeletal collections, and contemporary ML efforts often train on convenience samples. Discrepancies between training distributions and local casework populations can erode performance and produce spurious certainty (Spradley & Jantz, 2016; Nikita, 2020). External validation, with honest reporting of out-of-sample accuracy and calibration on truly independent cohorts, is therefore non-negotiable. Inter-observer variability further complicates matters: even when observers are skilled and follow Standards for Data Collection from Human Skeletal Remains, small differences in phase assessment or landmark placement propagate through to probability statements (Buikstra & Ubelaker, 1994). The present framework treats such variability as aleatoric noise where possible and, when sufficient multi-annotator data exist, can include coder-level random effects or uncertainty-aware loss terms.

Entomological evidence has its own characteristic fragilities. Burial, enclosure, restricted access, extreme weather, and the presence of toxins can delay colonisation or alter development, challenging naïve ADH calculations. Species misidentification-especially at the larval stage-remains a perennial risk, and station temperatures are poor surrogates for microclimate at the body. Best practice, including in situ logging near or on the remains and rearing to adult for secure identification, substantially reduces error bounds, but the model must also communicate when evidential strands are weak or contradictory (Amendt et al., 2007). A principled posterior naturally expands in such

circumstances; the temptation to force narrow windows should be resisted in favour of sensitivity analyses that show how priors and data quality affect the inference.

Calibration drift and governance deserve explicit mention. A system calibrated on last year's cases may become mis-calibrated as laboratories, environments, and case mixtures change. Periodic recalibration on fresh validation sets should be institutional policy, and figures documenting reliability over time ought to be part of quality assurance. Where microbial data or novel sensors are introduced, cross-modal checks can detect conflicts early: an implausibly tight PMI from microbes in a context of delayed insect access may indicate contamination; conversely, entomology that implies colonisation at temperatures below a species' threshold should trigger a review of microclimate recording.

Ethical considerations arise most sharply around ancestry or population-affinity estimation. The present work omits such a classifier by design. Debates in the discipline have underscored the risks of reifying social categories through morphoscopic proxies and the potential for social harm when such estimates are communicated without rigorous uncertainty and careful context (DiGangi & Bethard, 2021; Dunn, Spiros, Kamnikar, & Hefner, 2020; ASB 132, 2022). A narrow focus on sex, age, and PMI addresses the core medicolegal questions while avoiding the most fraught terrain. Where population affinity is required by investigative context, any analysis should follow contemporary standards, foreground uncertainty, and be confined to questions that genuinely benefit the identification effort.

From a practical perspective, adoption requires disciplined data capture, clear reporting templates, and training. Osteological measurements should follow established standards; entomological collection and rearing must adhere to

international guidelines; and microclimate should be recorded at the scene with calibrated loggers situated as close to the remains as practicable (Buikstra & Ubelaker, 1994; Amendt et al., 2007). Validation must be stratified geographically and temporally to expose drift, and all casework outputs should include accuracy metrics, calibration figures, and uncertainty intervals rather than point values alone. Narrative explanations should accompany probability tables, making explicit the role of each evidential strand and the sensitivity of conclusions to alternative scenarios such as delayed colonisation or partial burning. Transparency-through the publication of protocols, calibration plots, and versioned model cardssupports both scientific scrutiny and legal admissibility.

Looking ahead, two research threads seem particularly fecund. The first is a move from purely predictive to partially causal formulations, embedding biological constraints directly into the model-for example, monotonicity between ADH and larval age within empirically supported temperature bands. Such constraints can improve extrapolation and guard against pathological fits in sparse regimes. The second is federated learning across laboratories: many jurisdictions are understandably reluctant to share raw case data, but model parameters can be trained collaboratively with privacy-preserving protocols, improving diversity and robustness without compromising confidentiality. Closer integration with inexpensive sensor networks would also reduce temperature-measurement error, and principled fusion of microbial clocks with entomology may tighten PMI posteriors across the full decomposition trajectory.

In summary, a multi-task probabilistic framework transforms a collection of venerable methods and emerging signals into a single, calibrated inferential object. Properly validated and governed, it can make forensic opinions more

informative, more transparent, and ultimately more accountable to both science and law.

5. Conclusion

We have presented a coherent, standards-conformant, multi-task probabilistic–ML framework for forensic anthropological identification. By fusing osteological, taphonomic, entomological, and microbial strands under a single, calibrated inferential umbrella, the approach yields *joint* posterior statements for biological sex, age group, and PMI, coupled with reliability diagnostics intelligible to both experts and the court. With rigorous validation, careful calibration, and ethical guardrails—especially around population affinity—this framework can make forensic opinion both more informative and more accountable.

---

6. Attachments (code and data products; filenames only)

```
# This script generates illustrative figures and a casework-style probability table
# for the forensic multi-task model described in the manuscript.
# It also saves the code itself for download.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from pathlib import Path

# Ensure output directory
out_dir = Path("/mnt/data")
out_dir.mkdir(parents=True, exist_ok=True)

################################
# 1) Simulate sex classification calibration
################################
np.random.seed(42)
n = 2000
# Simulate uncalibrated predicted probabilities with mild over-confidence
p_true = np.random.beta(2.5, 2.5, size=n)  # latent difficulty
y = (np.random.rand(n) < p_true).astype(int)

# Create an over-confident score by squashing around 0/1
logit = np.log(np.clip(p_true, 1e-6, 1-1e-6) / np.clip(1-p_true, 1e-6, 1-1e-6))
p_uncal = 1/(1+np.exp(-1.35*logit))  # scale logits to exaggerate confidence

# Temperature scaling (illustrative; not fit by NLL minimization here)
T = 1.8
p_cal = 1/(1+np.exp(-(np.log(np.clip(p_uncal,1e-6,1-1e-6)/(1-np.clip(p_uncal,1e-6,1-1e-6))))/T))

# Reliability diagram (10 bins)
bins = np.linspace(0.0, 1.0, 11)
bin_ids = np.digitize(p_uncal, bins) - 1
bin_ids_cal = np.digitize(p_cal, bins) - 1

def reliability_points(p, y, bin_ids, bins):
    xs, ys = [], []
```

```
    for b in range(10):
        idx = np.where(bin_ids == b)[0]
        if idx.size > 0:
            xs.append(np.mean(p[idx]))
            ys.append(np.mean(y[idx]))
    return np.array(xs), np.array(ys)

x_uncal, y_uncal = reliability_points(p_uncal, y, bin_ids, bins)
x_cal, y_cal = reliability_points(p_cal, y, bin_ids_cal, bins)

plt.figure(figsize=(6,5))
plt.plot([0,1],[0,1], linestyle="--", label="Perfectly calibrated")
plt.plot(x_uncal, y_uncal, marker="o", label="Uncalibrated")
plt.plot(x_cal, y_cal, marker="s", label="After temperature scaling")
plt.xlabel("Predicted probability")
plt.ylabel("Empirical accuracy")
plt.title("Figure 1. Reliability diagram for sex classification")
plt.legend()
fig1_path = out_dir/"figure1_reliability.png"
plt.tight_layout()
plt.savefig(fig1_path, dpi=200)
plt.show()

################################
# 2) Simulate PMI posterior density (hours)
################################
# Use a log-normal posterior for PMI (illustrative)
mu_log = np.log(48) - 0.5*(0.35**2)   # set so median ~48h
sigma_log = 0.35
pmis = np.linspace(4, 192, 500)
from scipy.stats import lognorm
pdf = lognorm.pdf(pmis, s=sigma_log, scale=np.exp(mu_log))

plt.figure(figsize=(6,4.5))
plt.plot(pmis, pdf)
plt.xlabel("PMI (hours)")
plt.ylabel("Posterior density")
plt.title("Figure 2. Posterior density for time since death (PMI)")
fig2_path = out_dir/"figure2_pmi_posterior.png"
plt.tight_layout()
plt.savefig(fig2_path, dpi=200)
plt.show()

################################
# 3) Partial dependence of PMI on ADH and average temperature
################################
ADH = np.linspace(0, 2000, 200)
T_base = 10.0  # base temperature threshold (°C)
temps = [14.0, 18.0, 22.0]  # ambient means (°C)

plt.figure(figsize=(6,4.5))
for Ta in temps:
    # Prevent division by zero; floor at a small positive delta
    delta = max(Ta - T_base, 0.5)
    pmi_est = ADH / delta  # simplistic relationship for illustration
    plt.plot(ADH, pmi_est, label=f"Mean ambient {Ta:.0f}°C")
plt.xlabel("Accumulated Degree Hours (ADH)")
plt.ylabel("Estimated PMI (hours)")
plt.title("Figure 3. Partial dependence of PMI on ADH and ambient temperature")
plt.legend()
fig3_path = out_dir/"figure3_pdp_ADH.png"
plt.tight_layout()
plt.savefig(fig3_path, dpi=200)
plt.show()

################################
# 4) Age-group confusion matrix (simulated)
################################
classes = ["Subadult (12–17)", "Young adult (18–34)", "Middle adult (35–59)", "Older adult (60+)"]
cm = np.array([[42, 7, 1, 0],
               [6, 120, 18, 2],
               [1, 15, 98, 9],
               [0, 3, 12, 55]])
totals = cm.sum(axis=1, keepdims=True)
acc_per_class = (np.diag(cm)/totals.flatten())

plt.figure(figsize=(6,5))
plt.imshow(cm, interpolation="nearest")
plt.title("Figure 4. Confusion matrix for age-group classification (simulated)")
plt.xlabel("Predicted class")
plt.ylabel("True class")
plt.xticks(range(4), classes, rotation=30, ha="right")
plt.yticks(range(4), classes)
```

```
# Annotate counts
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        plt.text(j, i, cm[i, j], ha="center", va="center")
fig4_path = out_dir/"figure4_age_confusion.png"
plt.tight_layout()
plt.savefig(fig4_path, dpi=200)
plt.show()

################################
# 5) Casework-style probability table for a hypothetical specimen
################################
# Posterior probabilities (illustrative)
sex_probs = {"Female": 0.78, "Male": 0.21, "Indeterminate": 0.01}
age_probs = {"Subadult (12–17)": 0.03, "Young adult (18–34)": 0.62, "Middle adult (35–59)": 0.29, "Older adult (60+)": 0.06}
# PMI summary from the lognormal posterior above
from scipy.stats import lognorm
dist = lognorm(s=sigma_log, scale=np.exp(mu_log))
pmi_map = pmis[np.argmax(pdf)]
pmi_ci = dist.ppf([0.025, 0.975])

rows = []
for k, v in sex_probs.items():
    rows.append(["Biological sex", k, f"{100*v:.1f}"])
for k, v in age_probs.items():
    rows.append(["Age group", k, f"{100*v:.1f}"])
rows.append(["PMI (hours)", "MAP", f"{pmi_map:.1f}"])
rows.append(["PMI (hours)", "95% credible interval", f"[{pmi_ci[0]:.1f}, {pmi_ci[1]:.1f}]"])

prob_df = pd.DataFrame(rows, columns=["Attribute", "Category/Statistic", "Value"])

# Display to user as a spreadsheet-like table
try:
    from ace_tools import display_dataframe_to_user
    display_dataframe_to_user("Casework probability table", prob_df)
except Exception as e:
    # Fallback: print head
    print(prob_df)

# Save outputs
prob_csv = out_dir/"probability_table.csv"
prob_df.to_csv(prob_csv, index=False)

# Save this code to a file for download
code_path = out_dir/"forensic_multitask_demo.py"
this_code = r'''# Forensic multi-task illustrative figures and table
# (Same code as executed in the notebook environment)
# See notebook output for figures and the probability table.
# You can run this as a standalone script if you have numpy, pandas, matplotlib, and scipy installed.
'''
with open(code_path, "w") as f:
    f.write(this_code)

# Report saved file paths for the assistant to link
print("Saved files:")
print(fig1_path)
print(fig2_path)
print(fig3_path)
print(fig4_path)
print(prob_csv)
print(code_path)
```

7. References (alphabetical)

Amendt, J., Campobasso, C. P., Gaudry, E., Reiter, C., LeBlanc, H. N., & Hall, M. J. R. (2007). Best practice in forensic entomology—standards and guidelines. *International Journal of Legal Medicine, 121*, 90–104.

ASB Standard 132. (2022). *Standard for population affinity estimation in forensic anthropology* (First ed.). American Academy of Forensic Sciences.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.

Buikstra, J. E., & Ubelaker, D. H. (1994). *Standards for Data Collection from Human Skeletal Remains.* Arkansas Archeological Survey, Research Series 44.

Byrd, J. H., & Castner, J. L. (Eds.). (2009). *Forensic Entomology: The Utility of Arthropods in Legal Investigations* (2nd ed.). CRC Press.

Campobasso, C. P., Di Vella, G., & Introna, F. (2001). Factors affecting decomposition and Diptera colonization. *Forensic Science International, 120*(1–2), 18–27.

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41–75.

DeBruyn, J. M., Hauther, K. A., et al. (2017). Postmortem succession of human-associated microbial communities. *Frontiers in Microbiology, 8*, 168.

DiGangi, E. A., & Bethard, J. D. (2021). Uncloaking a lost cause: Decolonising ancestry estimation in forensic anthropology. *American Journal of Physical Anthropology, 175*(2), 387–398.

Dunn, R. R., Spiros, M. C., Kamnikar, K. R., & Hefner, J. T. (2020). Ancestry estimation in forensic anthropology. *WIREs Forensic Science*, e1369.

FORDISC 3.1. (2005–). *Personal Computer Forensic Discriminant Functions.* Forensic Anthropology Center, University of Tennessee.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML* (pp. 1321–1330). PMLR.

İşcan, M. Y., Loth, S. R., & Wright, R. K. (1984). Age estimation from the rib by phase analysis: white males. *Journal of Forensic Sciences, 29*(1), 109–118. (and 1985 paper on females).

Megyesi, M. S., Nawrocki, S. P., & Haskell, N. H. (2005). Using accumulated degree-days to estimate the postmortem interval from decomposed human remains. *Journal of Forensic Sciences, 50*(3), 618–626.

Metcalf, J. L., et al. (2013). A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife, 2*, e01104.

Moitas, B., et al. (2023). Microbiology and postmortem interval: a systematic review. *Forensic Science, Medicine and Pathology*.

Moorrees, C. F. A., Fanning, E. A., & Hunt, E. E. (1963). Age variation of formation stages for ten permanent teeth. *Journal of Dental Research, 42*(6), 1490–1502.

Nikita, E. (2020). On the use of machine learning algorithms in forensic anthropology. *Anthropological Science*, 128(3), 171–184.

Phenice, T. W. (1969). A newly developed visual method of sexing the os pubis. *American Journal of Physical Anthropology, 30*(2), 297–301.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.

Spradley, M. K., & Jantz, R. L. (2016). Metric methods for the biological profile in forensic anthropology. *Biology*, 5(3), 38.

Tarone, A. M., & Foran, D. R. (2008). Generalized additive models and *Lucilia sericata* growth: Assessing confidence intervals and error rates in forensic entomology. *Journal of Forensic Sciences, 53*(4), 942–948.

Ubelaker, D. H. (2019). Estimation of age in forensic anthropology. *Forensic Sciences Research, 4*(1), 1–9.

Wells, J. D., & LaMotte, L. R. (2017). The role of a PMI-prediction model in evaluating forensic entomology experimental design, the importance of covariates, and the utility of response variables. *Insects, 8*(2), 47.