Online Appendix for "Adversarial Training-Based Deep Imbalanced Learning"

Appendix A: Proof of Theorem 1 and Related Lemmas

A.1. Proof of Lemma 1

LEMMA 1. *Let p be the probability density function of $\mathbb{D}_+$ with support $\mathbb{R}^d$. For $\sigma \to 0$, we have:*

$$L_{AAE} = \mathbb{E}[\|x - r(x)\|_2^2] + \sigma^2 \mathbb{E}[\| \tfrac{\partial r(x)}{\partial x} \|_F^2] + \lambda \mathbb{E}[f(r(x))] + o(\sigma^2). \tag{A1}$$

Proof. By Taylor expansion, $r(x+\epsilon) = r(x) + \frac{\partial r(x)}{\partial x}\epsilon + o(\sigma^2)$. Substituting this into L_{AAE}:

$$\begin{aligned} L_{\mathrm{AAE}} &= \mathbb{E}\Bigg[\left(r(x)^T - x^T + \epsilon^T \tfrac{\partial r(x)}{\partial x}^T \right) \left(r(x) - x + \tfrac{\partial r(x)}{\partial x}\epsilon \right) + \lambda f(r(x)) + \lambda \tfrac{\partial f(r(x))}{\partial r(x)} \tfrac{\partial r(x)}{\partial x}\epsilon \Bigg] + o(\sigma^2) \\ &= \mathbb{E}\Bigg[(r(x)^T - x^T)(r(x) - x) + (r(x)^T - x^T)\tfrac{\partial r(x)}{\partial x}\epsilon + \epsilon^T \tfrac{\partial r(x)}{\partial x}^T (r(x) - x) \\ &\quad + \epsilon^T \tfrac{\partial r(x)}{\partial x}^T \tfrac{\partial r(x)}{\partial x}\epsilon + \lambda f(r(x)) + \lambda \tfrac{\partial f(r(x))}{\partial r(x)} \tfrac{\partial r(x)}{\partial x}\epsilon \Bigg] + o(\sigma^2). \end{aligned} \tag{A2}$$

Since ϵ and x are independent and $\mathbb{E}[\epsilon] = 0$, the linear terms in ϵ vanish:

$$L_{\mathrm{AAE}} = \mathbb{E}[\|r(x) - x\|_2^2] + \mathbb{E}[\epsilon^T \tfrac{\partial r(x)}{\partial x}^T \tfrac{\partial r(x)}{\partial x}\epsilon] + \lambda \mathbb{E}[f(r(x))] + o(\sigma^2). \tag{A3}$$

Using the trace trick for the quadratic term $\epsilon^T \frac{\partial r(x)}{\partial x}^T \frac{\partial r(x)}{\partial x}\epsilon$ and $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I$:

$$\mathbb{E}[\epsilon^T \tfrac{\partial r(x)}{\partial x}^T \tfrac{\partial r(x)}{\partial x}\epsilon] = \mathrm{Tr}\left(\mathbb{E}\left[\epsilon\epsilon^T \tfrac{\partial r(x)}{\partial x}^T \tfrac{\partial r(x)}{\partial x} \right] \right) = \sigma^2 \mathbb{E}\left[\| \tfrac{\partial r(x)}{\partial x} \|_F^2 \right]. \tag{A4}$$

Substituting Equation (A4) into Equation (A3) yields the final result:

$$L_{\mathrm{AAE}} = \mathbb{E}[\|r(x) - x\|_2^2] + \sigma^2 \mathbb{E}[\| \tfrac{\partial r(x)}{\partial x} \|_F^2] + \lambda \mathbb{E}[f(r(x))] + o(\sigma^2). \tag{A5}$$

This completes the proof.

A.2. Proof of Theorem 1

THEOREM 1. *Given:*

$$\begin{aligned} L_{\mathrm{AAE}} &= \mathbb{E}\left[\|r(x) - x\|_2^2 \right] + \sigma^2 \mathbb{E}\left[\left\| \frac{\partial r(x)}{\partial x} \right\|_F^2 \right] + \lambda \mathbb{E}\left[f\left(r(x) \right) \right] \\ &= \int_{x \in \mathbb{R}^d} p(x) \left[\|r(x) - x\|_2^2 + \sigma^2 \left\| \frac{\partial r(x)}{\partial x} \right\|_F^2 + \lambda f\left(r(x) \right) \right] dx, \end{aligned}$$

where $p(x)$ is first-order differentiable, and $r : \mathbb{R}^d \to \mathbb{R}^d$ is second-order differentiable. Let $r^*(x)$ represent the optimal AAE that minimizes L_{AAE}, then:

$$\begin{aligned} r^*(x) = x &- \frac{\lambda}{2} \frac{\partial f\left(r(x) \right)}{\partial r(x)} + \sigma^2 \frac{\partial \log p(x)}{\partial x} \\ &- \frac{\lambda \sigma^2}{2} \sum_{i=1}^{d} \left(\frac{\partial^3 f\left(r(x) \right)}{\partial x_i^2 \partial r(x)} + \frac{\partial \log p(x)}{\partial x_i} \frac{\partial^2 f\left(r(x) \right)}{\partial x_i \partial r(x)} \right) + o\left(\sigma^2 \right). \end{aligned} \tag{A6}$$

Proof. Let

$$F\left(x_1, x_2, \ldots, x_d, r, r_{x_1}, r_{x_2}, \ldots, r_{x_d}\right) = p(x)\left[\left\|r(x) - x\right\|_2^2 + \sigma^2\left\|\frac{\partial r(x)}{\partial x}\right\|_F^2 + \lambda f\left(r(x)\right)\right]$$

represent a functional defined over the function r, where

$$x = \left(x_1, x_2, \ldots, x_d\right), r(x) = \left(r_1(x), r_2(x), \ldots, r_d(x)\right), r_{x_i} = \frac{\partial F}{\partial x_i}$$

We express L_{AAE} in a more concrete form:

$$L_{\text{AAE}} = \int_{\mathbb{R}^d} p(x)\left[\sum_{i=1}^{d}\left(r_i(x) - x_i\right)^2 + \sigma^2\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial r_i(x)}{\partial x_j}^2 + \lambda f\left(r(x)\right)\right]dx. \tag{A7}$$

According to the Euler-Lagrange equation, when L_{AAE} attains its minimum, r should satisfy:

$$\frac{\partial F}{\partial r} = \sum_{i=1}^{d}\frac{\partial}{\partial x_i}\frac{\partial F}{\partial r_{x_i}}. \tag{A8}$$

We have:

$$\frac{\partial F}{\partial r} = p(x)\left(2r(x) - 2x + \lambda\frac{\partial f\left(r(x)\right)}{\partial r(x)}\right), \tag{A9}$$

$$\frac{\partial F}{\partial r_{x_i}} = 2\sigma^2 p(x)\left[\frac{\partial r_1(x)}{\partial x_i}, \frac{\partial r_2(x)}{\partial x_i}, \ldots, \frac{\partial r_d(x)}{\partial x_i}\right]^T, \tag{A10}$$

$$\begin{aligned}\frac{\partial}{\partial x_i}\left(\frac{\partial F}{\partial r_{x_i}}\right) = {} & 2\sigma^2\frac{\partial p(x)}{\partial x_i}\left[\frac{\partial r_1(x)}{\partial x_i}, \frac{\partial r_2(x)}{\partial x_i}, \ldots, \frac{\partial r_d(x)}{\partial x_i}\right]^T \\ & + 2\sigma^2 p(x)\left[\frac{\partial^2 r_1(x)}{\partial x_i^2}, \frac{\partial^2 r_2(x)}{\partial x_i^2}, \ldots, \frac{\partial^2 r_d(x)}{\partial x_i^2}\right]^T.\end{aligned} \tag{A11}$$

Therefore, Equation (A8) transforms into:

$$p(x)\left(2r(x) - 2x + \lambda\frac{\partial f\left(r(x)\right)}{\partial r(x)}\right) = 2\sigma^2\sum_{i=1}^{d}\begin{bmatrix}\frac{\partial p(x)}{\partial x_i}\frac{\partial r_1(x)}{\partial x_i} + p(x)\frac{\partial^2 r_1(x)}{\partial x_i^2} \\ \vdots \\ \frac{\partial p(x)}{\partial x_i}\frac{\partial r_d(x)}{\partial x_i} + p(x)\frac{\partial^2 r_d(x)}{\partial x_i^2}\end{bmatrix}. \tag{A12}$$

Equation (A12) can be decomposed into d equations, and for $k = 1, \cdots, d$, we have:

$$p(x)\left(r_k(x) - x_k + \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)}\right) = \sigma^2\sum_{i=1}^{d}\left(\frac{\partial p(x)}{\partial x_i}\frac{\partial r_k(x)}{\partial x_i} + p(x)\frac{\partial^2 r_k(x)}{\partial x_i^2}\right). \tag{A13}$$

Furthermore, according to the assumption that $p(x) \neq 0$, dividing both sides of Equation (A13) by $p(x)$ yields:

$$r_k(x) - x_k + \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} = \sigma^2\sum_{i=1}^{d}\left(\frac{\partial log p(x)}{\partial x_i}\frac{\partial r_k(x)}{\partial x_i} + \frac{\partial^2 r_k(x)}{\partial x_i^2}\right). \tag{A14}$$

It is not difficult to observe that the partial differential equation in Equation (A14) is a recursive relation concerning $r_k(x)$, therefore:

$$\begin{aligned}r_k(x) &= x_k - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} + \sigma^2\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial r_k(x)}{\partial x_i} + \frac{\partial^2 r_k(x)}{\partial x_i^2}\right) \\ &= x_k - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} + \sigma^2\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}I(i=k)\right) + \sigma^2\sum_{i=1}^{d}\frac{\partial^2 r_k(x)}{\partial x_i^2}\end{aligned}$$

$$
\begin{aligned}
&- \frac{\lambda\sigma^2}{2}\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial}{\partial x_i}\left(\frac{\partial f\left(r(x)\right)}{\partial r_k(x)}\right)\right) \\
&+ \sigma^{2^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial}{\partial x_i}\left(\frac{\partial \log p(x)}{\partial x_j}\frac{\partial r_k(x)}{\partial x_j}+\frac{\partial^2 r_k(x)}{\partial x_j^2}\right)\right) \\
&= x_k - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} + \sigma^2\frac{\partial \log p(x)}{\partial x_k} + \sigma^2\sum_{i=1}^{d}\frac{\partial^2 r_k(x)}{\partial x_i^2} \\
&- \frac{\lambda\sigma^2}{2}\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r_k(x)}\right) + \sigma^{2^2}\rho\left(\sigma^2, x\right). \qquad \text{(A15)}
\end{aligned}
$$

In fact, $\sum_{i=1}^{d}\frac{\partial^2 r_k(x)}{\partial x_i^2}$ contains terms with quadratic or higher powers of σ; thus, Equation (A15) can be further simplified. To achieve this, let's consider taking the second partial derivative of r with respect to a certain x_l:

$$
\begin{aligned}
\frac{\partial r_k(x)}{\partial x_l} &= I\left(l=k\right) - \frac{\lambda}{2}\frac{\partial^2 f\left(r(x)\right)}{\partial x_l \partial r_k(x)} + \sigma^2\frac{\partial^2 \log p(x)}{\partial x_l \partial x_k} \\
&+ \sigma^2\frac{\partial}{\partial x_l}\left(\sum_{i=1}^{d}\frac{\partial^2 r_k(x)}{\partial x_i^2} - \frac{\lambda}{2}\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r_k(x)}\right) + \sigma^2\rho\left(\sigma^2, x\right)\right), \qquad \text{(A16)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 r_k(x)}{\partial x_l^2} &= -\frac{\lambda}{2}\frac{\partial^3 f\left(r(x)\right)}{\partial x_l^2 \partial r_k(x)} + \sigma^2\frac{\partial^3 \log p(x)}{\partial x_l^2 \partial x_k} \\
&+ \sigma^2\frac{\partial}{\partial x_l^2}\left(\sum_{i=1}^{d}\frac{\partial^2 r_k(x)}{\partial x_i^2} - \frac{\lambda}{2}\sum_{i=1}^{d}\left(\frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r_k(x)}\right) + \sigma^2\rho\left(\sigma^2, x\right)\right). \qquad \text{(A17)}
\end{aligned}
$$

Substituting Equations (A16) and (A17) into Equation (A12), we have:

$$
\begin{aligned}
r_k(x) &= x_k - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} + \sigma^2\frac{\partial \log p(x)}{\partial x_k} \\
&- \frac{\lambda\sigma^2}{2}\sum_{i=1}^{d}\left(\frac{\partial^3 f\left(r(x)\right)}{\partial x_i^2 \partial r_k(x)} + \frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r_k(x)}\right) + \sigma^{2^2}\eta\left(\sigma^2, x\right) \qquad \text{(A18)} \\
&= x_k - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r_k(x)} + \sigma^2\frac{\partial \log p(x)}{\partial x_k} \\
&- \frac{\lambda\sigma^2}{2}\sum_{i=1}^{d}\left(\frac{\partial^3 f\left(r(x)\right)}{\partial x_i^2 \partial r_k(x)} + \frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r_k(x)}\right) + \sigma^{2^2}\eta\left(\sigma^2, x\right). \qquad \text{(A19)}
\end{aligned}
$$

Then, we have:

$$
\begin{aligned}
r(x) &= x - \frac{\lambda}{2}\frac{\partial f\left(r(x)\right)}{\partial r(x)} + \sigma^2\frac{\partial \log p(x)}{\partial x} \\
&- \frac{\lambda\sigma^2}{2}\sum_{i=1}^{d}\left(\frac{\partial^3 f\left(r(x)\right)}{\partial x_i^2 \partial r(x)} + \frac{\partial \log p(x)}{\partial x_i}\frac{\partial^2 f\left(r(x)\right)}{\partial x_i \partial r(x)}\right) + o\left(\sigma^2\right). \qquad \text{(A20)}
\end{aligned}
$$

This completes the proof of Theorem 1.

Appendix B: Proof of Theorem 2 and Related Lemmas

B.1. Proof of Lemma 2

LEMMA 2. *Based on Assumptions 2 and 3, $L_s(\theta)$ is L-smooth where $L = L_{\theta\phi}L_{\phi\theta}/\mu + L_{\theta\theta}$, i.e., for any θ_1 and θ_2, the following holds:*

$$
\left\|\nabla L_s\left(\theta_1\right) - \nabla L_s\left(\theta_2\right)\right\|_2 \leq L\left\|\theta_1 - \theta_2\right\|_2, \qquad \text{(A21)}
$$

$$
L_s\left(\theta_1\right) \leq L_s\left(\theta_2\right) + \left\langle \nabla L_s\left(\theta_2\right), \theta_1 - \theta_2\right\rangle + \frac{L}{2}\left\|\theta_1 - \theta_2\right\|_2^2. \qquad \text{(A22)}
$$

Proof. According to Assumption 2, for any given θ_1 and θ_2, we have:

$$h\left(\theta_2, \phi^*\left(\theta_1\right), x\right) \leq h\left(\theta_2, \phi^*\left(\theta_2\right), x\right) + \left\langle \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_2\right), x\right), \phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\rangle - \frac{\mu}{2}\left\|\phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\|_2^2. \tag{A23}$$

This inequality holds because $\left\langle \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_2\right), x\right), \phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\rangle \leq 0$. Similarly, according to Assumption 2, we know that:

$$h\left(\theta_2, \phi^*\left(\theta_2\right), x\right) \leq h\left(\theta_2, \phi^*\left(\theta_1\right), x\right) + \left\langle \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_1\right), x\right), \phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\rangle - \frac{\mu}{2}\left\|\phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\|_2^2. \tag{A24}$$

Combining inequalities (A23) and (A24) yields:

$$\begin{aligned}
\mu\left\|\phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\|_2^2 &\leq \left\langle \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_1\right), x\right), \phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\rangle \\
&\leq \left\langle \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_1\right), x\right) - \nabla_\phi h\left(\theta_1, \phi^*\left(\theta_1\right), x\right), \phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\rangle \\
&\leq \left\|\nabla_\phi h\left(\theta_2, \phi^*\left(\theta_1\right), x\right) - \nabla_\phi h\left(\theta_1, \phi^*\left(\theta_1\right), x\right)\right\|_2 \left\|\phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\|_2 \\
&\leq L_{\phi\theta}\left\|\theta_2 - \theta_1\right\|_2 \left\|\phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\|,
\end{aligned} \tag{A25}$$

where the second inequality holds because $\left\langle \nabla_\phi h\left(\theta_1, \phi^*\left(\theta_1\right), x\right), \phi^*\left(\theta_2\right) - \phi^*\left(\theta_1\right)\right\rangle \leq 0$, the third uses Cauchy-Schwarz inequality, and the last uses Assumption 1. Inequality (A25) finally becomes:

$$\left\|\phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\|_2 \leq \frac{L_{\phi\theta}}{\mu}\left\|\theta_2 - \theta_1\right\|_2. \tag{A26}$$

Next, we have:

$$\begin{aligned}
&\left\|\nabla_\phi h\left(\theta_1, \phi^*\left(\theta_1\right), x\right) - \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_2\right), x\right)\right\|_2 \\
&\leq \left\|\nabla_\phi h\left(\theta_1, \phi^*\left(\theta_1\right), x\right) - \nabla_\phi h\left(\theta_1, \phi^*\left(\theta_2\right), x\right)\right\|_2 \\
&\quad + \left\|\nabla_\phi h\left(\theta_1, \phi^*\left(\theta_2\right), x\right) - \nabla_\phi h\left(\theta_2, \phi^*\left(\theta_2\right), x\right)\right\|_2 \\
&\leq L_{\theta\phi}\left\|\phi^*\left(\theta_1\right) - \phi^*\left(\theta_2\right)\right\|_2 + L_{\theta\theta}\left\|\theta_1 - \theta_2\right\|_2 \\
&\leq \left(\frac{L_{\theta\phi}L_{\phi\theta}}{\mu} + L_{\theta\theta}\right)\left\|\theta_1 - \theta_2\right\|_2,
\end{aligned} \tag{A27}$$

where the first inequality follows from the triangle inequality, the second uses Assumption 1, and the last uses inequality (A26). Finally, by the definition of $L_s(\theta)$, we have:

$$\begin{aligned}
&\left\|\nabla L_s\left(\theta_1\right) - \nabla L_s\left(\theta_2\right)\right\|_2 \\
&= \left\|\frac{1}{N + N_{\text{adv}}}\left(\sum_{i=1}^{N+N_{\text{adv}}} \nabla_\phi h\left(\theta, \phi^*\left(\theta_1\right), x_i\right) - \sum_{i=1}^{N+N_{\text{adv}}} \nabla_\phi h\left(\theta, \phi^*\left(\theta_2\right), x_i\right)\right)\right\|_2 \\
&\leq \frac{1}{N + N_{\text{adv}}}\sum_{i=1}^{N+N_{\text{adv}}}\left\|\nabla_\phi h\left(\theta, \phi^*\left(\theta_1\right), x_i\right) - \nabla_\phi h\left(\theta, \phi^*\left(\theta_2\right), x_i\right)\right\| \\
&\leq \left(\left(\frac{L_{\theta\phi}L_{\phi\theta}}{\mu} + L_{\theta\theta}\right)\left\|\theta_1 - \theta_2\right\|_2\right).
\end{aligned} \tag{A28}$$

This completes the proof of Lemma 2.

B.2. Proof of Lemma 3

LEMMA 3. *Under Assumptions 1 and 2, the approximate stochastic gradient $\hat{g}(\theta)$ satisfies:*

$$\|\hat{g}(\theta) - g(\theta)\|_2 \leq L_{\theta\phi}\sqrt{\frac{\delta}{\mu}}. \tag{A29}$$

Proof. By the triangle inequality and Assumption 1, we have:

$$\begin{aligned}\|\hat{g}(\theta) - g(\theta)\|_2 &= \left\| \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} \left(\nabla_\theta h(\theta, \hat{\phi}(\theta), x) - \nabla_\theta h(\theta, \phi^*(\theta), x)\right)\right\|_2 \\ &\leq \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} L_{\theta\phi}\left\|\hat{\phi}(\theta) - \phi^*(\theta)\right\|_2 = L_{\theta\phi}\left\|\hat{\phi}(\theta) - \phi^*(\theta)\right\|_2.\end{aligned} \tag{A30}$$

Based on the μ-strong concavity in Assumption 2, we have:

$$h(\theta, \hat{\phi}(\theta), x) \leq h(\theta, \phi^*(\theta), x) + \langle \nabla_\phi h(\theta, \phi^*(\theta), x), \hat{\phi}(\theta) - \phi^*(\theta)\rangle - \frac{\mu}{2}\|\hat{\phi}(\theta) - \phi^*(\theta)\|_2^2, \tag{A31}$$

$$h(\theta, \phi^*(\theta), x) \leq h(\theta, \hat{\phi}(\theta), x) + \langle \nabla_\phi h(\theta, \hat{\phi}(\theta), x), \phi^*(\theta) - \hat{\phi}(\theta)\rangle - \frac{\mu}{2}\|\phi^*(\theta) - \hat{\phi}(\theta)\|_2^2. \tag{A32}$$

Adding inequalities (A31) and (A32) yields:

$$\mu\|\phi^*(\theta) - \hat{\phi}(\theta)\|_2^2 \leq \langle \nabla_\phi h(\theta, \phi^*(\theta), x) - \nabla_\phi h(\theta, \hat{\phi}(\theta), x), \hat{\phi}(\theta) - \phi^*(\theta)\rangle. \tag{A33}$$

Since $\hat{\phi}(\theta)$ is a δ-approximation and $\phi^*(\theta)$ is the optimum, we have:

$$\langle \phi^*(\theta) - \hat{\phi}(\theta), \nabla_\phi h(\theta, \hat{\phi}(\theta), x)\rangle \leq \delta \quad \text{and} \quad \langle \hat{\phi}(\theta) - \phi^*(\theta), \nabla_\phi h(\theta, \phi^*(\theta), x)\rangle \leq 0. \tag{A34}$$

Summing these gives $\langle \hat{\phi}(\theta) - \phi^*(\theta), \nabla_\phi h(\theta, \phi^*(\theta), x) - \nabla_\phi h(\theta, \hat{\phi}(\theta), x)\rangle \leq \delta$. Thus, we obtain $\|\phi^*(\theta) - \hat{\phi}(\theta)\|_2 \leq \sqrt{\delta/\mu}$. Substituting this into inequality (A30) completes the proof.

B.3. Proof of Theorem 2

THEOREM 2. *Let $\Delta = L_s(\theta^0) - \min_\theta L_s(\theta)$, and $\eta_t = \eta = \sqrt{\Delta/(TL\vartheta^2)}$ represent the step size for the outer optimization in Problem (5). Based on Lemmas 2, 3, and Assumption 3, we have:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla L_s(\theta^t)\|_2^2\right] \leq 3\vartheta\sqrt{\frac{L\Delta}{T}} + \frac{L_{\theta\phi}^2\delta}{\mu}. \tag{A35}$$

Proof. According to Lemma 2, we have:

$$\begin{aligned}L_s(\theta^{t+1}) &\leq L_s(\theta^t) + \langle \nabla L_s(\theta^t), \theta^{t+1} - \theta^t\rangle + \frac{L}{2}\|\theta^{t+1} - \theta^t\|_2^2 \\ &= L_s(\theta^t) - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla L_s(\theta^t)\|_2^2 + \eta_t(1 - L\eta_t)\langle \nabla L_s(\theta^t), g(\theta^t) - \hat{g}(\theta^t)\rangle \\ &\quad + \eta_t(1 - L\eta_t)\langle \nabla L_s(\theta^t), \nabla L_s(\theta^t) - g(\theta^t)\rangle + \frac{L\eta_t^2}{2}\|\hat{g}(\theta^t) - g(\theta^t) + g(\theta^t) - \nabla L_s(\theta^t)\|_2^2 \\ &\leq L_s(\theta^t) - \frac{\eta_t}{2}\|\nabla L_s(\theta^t)\|_2^2 + \frac{\eta_t}{2}\|\hat{g}(\theta^t) - g(\theta^t)\|_2^2 \\ &\quad + \eta_t(1 - L\eta_t)\langle \nabla L_s(\theta^t), \nabla L_s(\theta^t) - g(\theta^t)\rangle + \frac{L\eta_t^2}{2}\|g(\theta^t) - \nabla L_s(\theta^t)\|_2^2,\end{aligned}$$

where the last inequality is due to Young's inequality. Taking the expectation on both sides of the equation with respect to the variable θ^t, we have:

$$\mathbb{E}\left[L_s(\theta^{t+1}) - L_s(\theta^t)\right] \leq -\frac{\eta_t}{2}\mathbb{E}\left[\|\nabla L_s(\theta^t)\|_2^2\right] + \frac{\eta_t L_{\theta\phi}^2\delta}{2\mu} + \frac{L\eta_t^2\vartheta^2}{2}. \tag{A36}$$

The inequality relies on Assumption 3 and Lemma 3. The summation is performed over the inequalities (A36) for $t = 0, 1, \cdots T-1$, resulting in:

$$\sum_{t=0}^{T-1} \frac{\eta_t}{2} \mathbb{E}\left[\left\|\nabla L_s\left(\theta^t\right)\right\|_2^2\right] \leq \mathbb{E}\left[L_s\left(\theta^0\right) - L_s\left(\theta^T\right)\right] + \sum_{t=0}^{T-1} \frac{\eta_t L_{\theta\phi}^2 \delta}{2\mu} + \sum_{t=0}^{T-1} \frac{L\eta_t^2 \vartheta^2}{2}. \tag{A37}$$

Considering that $\Delta = L_s\left(\theta^0\right) - \min_{\theta} L_s(\theta)$, we have:

$$\mathbb{E}\left[L_s\left(\theta^0\right) - L_s\left(\theta^T\right)\right] \leq \Delta. \tag{A38}$$

Substituting inequality (A38) and $\eta_t = \sqrt{\Delta/\left(TL\vartheta^2\right)}$ into inequality (A37), we obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla L_s\left(\theta^t\right)\right\|_2^2\right] \leq 3\vartheta\sqrt{\frac{L\Delta}{T}} + \frac{L_{\theta\phi}^2 \delta}{\mu}. \tag{A39}$$

This completes the proof of Theorem 2.

Appendix C: Detailed Experimental Setup

To ensure robust evaluation, we employ ten runs of five-fold cross-validation. In each fold, the training data is further split into a 90% training subset and a 10% validation subset for hyperparameter optimization. The classifier is then retrained on the complete training set using the optimal parameters and evaluated on the held-out test set. The final results are reported as the average across all ten repetitions.

We employ a fixed DNN backbone for each dataset to control for model capacity, with architectures detailed in Table A1. The backbone comprises a 2–3 layer feedforward network (with Batch Normalization, ReLU, and Dropout) and a sigmoid output. Training uses dataset-adjusted batch sizes and runs for up to 200 epochs with early stopping (patience = 10). All imbalanced learning methods are evaluated on this standardized setup, ensuring that observed performance gaps result exclusively from the methods themselves.

As for the learning rate η of ATDIL, we assume $\eta = \sqrt{\Delta/(TL\vartheta^2)}$ to ensure convergence to a first-order stationary point. In practice, however, the constants L, ϑ, and Δ are generally unknown or difficult to estimate precisely. Therefore, we do not explicitly set the learning rate according to this theoretical formula. Instead, the empirically chosen value (1×10^{-3}) has been found to implicitly satisfy the theoretical stability condition, in the sense that the training loss decreases monotonically and the gradients remain bounded throughout the optimization process. The optimizer is uniformly set to Adam with this learning rate, which provides stable training dynamics and aligns with the theoretical stability condition discussed in Theorem 2.

Table A1 Backbone DNN Configurations per Dataset (Fixed Across All Compared Methods)

Dataset	Hidden layers (activation)	Dropout	Norm	Batch size	Learning rate
Home	[64 (ReLU), 32 (ReLU)]	0.1 / 0.1	BN	32	1×10^{-3}
LC	[128 (ReLU), 64 (ReLU), 32 (ReLU)]	0.1 / 0.1 / 0.1	BN	64	1×10^{-3}
IEEE	[256 (ReLU), 128 (ReLU), 64 (ReLU)]	0.1 / 0.1 / 0.1	BN	64	1×10^{-3}
Prosper	[64 (ReLU), 32 (ReLU)]	0.1 / 0.1	BN	32	1×10^{-3}
Bank	[128 (ReLU), 64 (ReLU), 32 (ReLU)]	0.1 / 0.1 / 0.1	BN	32	1×10^{-3}
CCFD-A	[64 (ReLU), 32 (ReLU)]	0.1 / 0.1	BN	128	1×10^{-3}
CCFD-B	[64 (ReLU), 32 (ReLU)]	0.1 / 0.1	BN	128	1×10^{-3}

Notes: (i) BN denotes Batch Normalization applied before each hidden layer's activation; (ii) The output layer uses a single sigmoid unit for binary classification; (iii) Early stopping on validation AUROC with patience $= 10$, maximum epochs $= 200$.

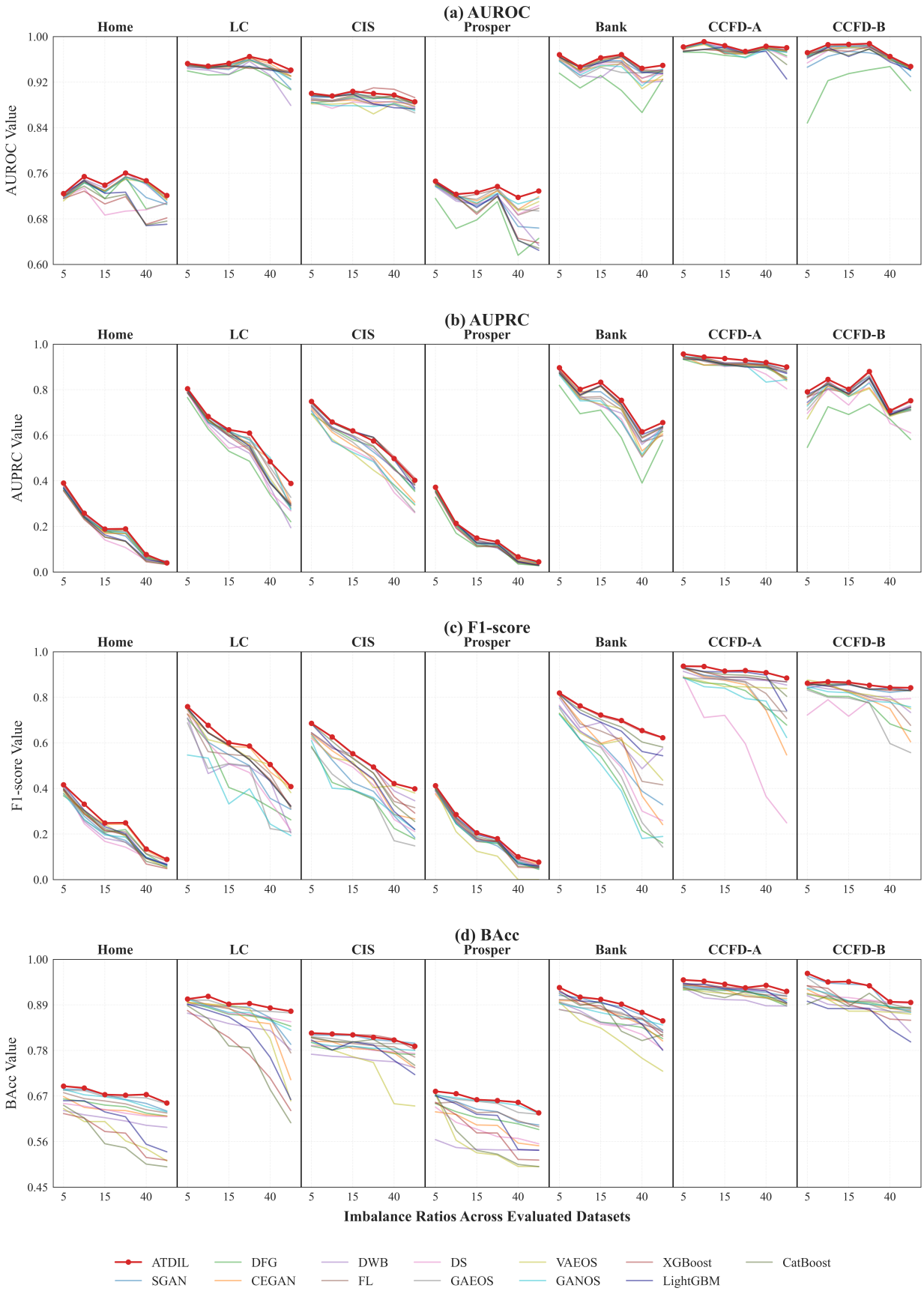

Figure A1 **Performance Comparison of Various Models Under Different Imbalance Ratios Across Seven Financial Fraud Datasets**

According to Algorithm 1, ATDIL requires configuring the AAE structure and parameters λ, σ, c, T. We implement AAE with four hidden layers, a squeeze-and-excitation module, and ReLU activation. Since the misclassification cost c has the most significant impact on performance, we fix $\lambda = 0.001$, $\sigma = 0.01$, and $T = 100$ to demonstrate robustness without extensive tuning. The optimal c is determined via a grid search on each dataset within the range $[5, 60]$ with a step size of 5.

The imbalanced learning methods SGAN, DFG, CEGAN, DS, GANOS and VAEOS are all based on the generative model VAE or GAN to synthesize the minority-class samples. We first determine the appropriate structure and parameters for the generative models involved in these methods, and then determine the IR of each dataset after resampling. Similarly, GAEOS parameters are set based on the recommendations in Zhang et al. (2022). For FL, we conduct a grid search for the focusing parameter $\gamma \in \{1, 2, 3\}$ and weighting factor $\alpha \in [0.1, 1]$ (step 0.1). DWB has a single hyperparameter, the misclassification cost, which is tuned using the same strategy as ATDIL.

Appendix D: Additional Experimental Results for Section 4.4 in the Main Text

This section presents additional experimental results that are not included in Section 4.4 of the main text. The detailed results are provided in Figure A1.

Appendix E: Additional Experimental Results for Section 4.5 in the Main Text

Table A2 presents ablation results, while Figure A2 shows that integrating ATDIL with tree-based ensembles (XGBoost, LightGBM, and CatBoost) yields consistent gains across various imbalance ratios.

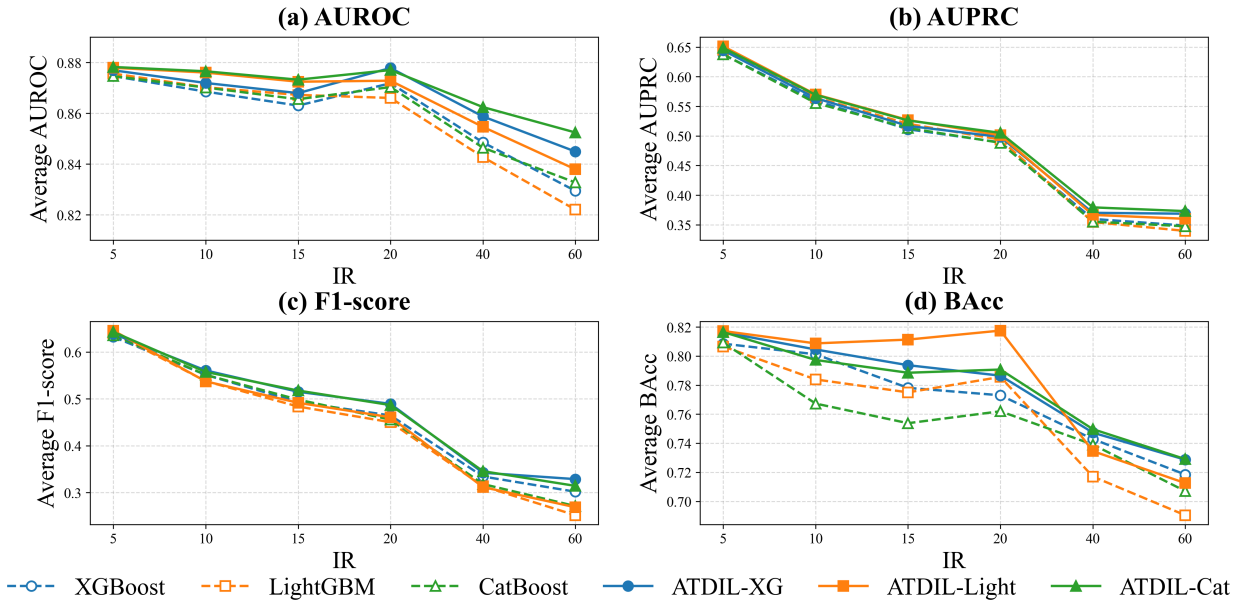

Figure A2 Performance Comparison of Tree-Based Models and Their ATDIL-Enhanced Versions Under Varying Imbalance Ratios

Appendix F: Additional Experimental Results for Section 4.7 in the Main Text

Figures A3-A4 provide additional experimental evidence supporting the security of our proposed approach.

References

Zhang J, Zhang L, Li G, Wu C (2022) Adversarial examples for good: Adversarial examples guided imbalanced learning. *Proc. 29th IEEE Internat. Conf. Image Process.*, 136–140 (Bordeaux: IEEE).

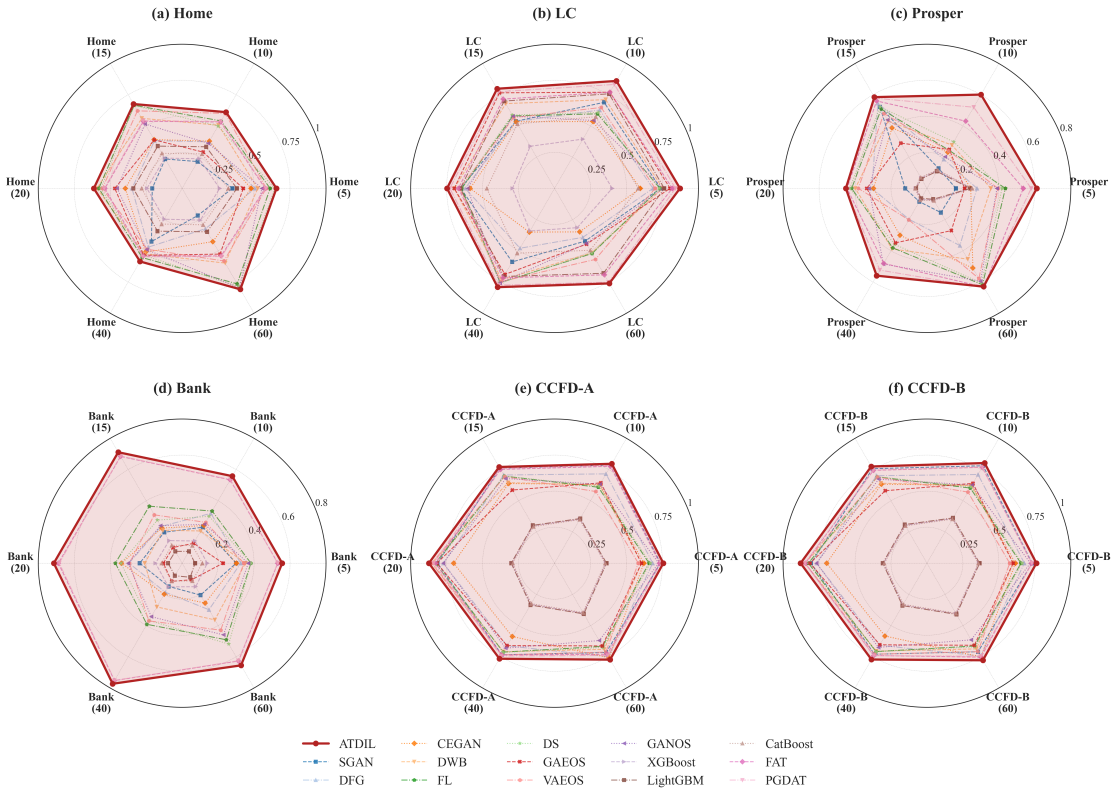

Figure A3 **Radar Visualization of SMS Values for Various Detection Models Across Six Datasets**

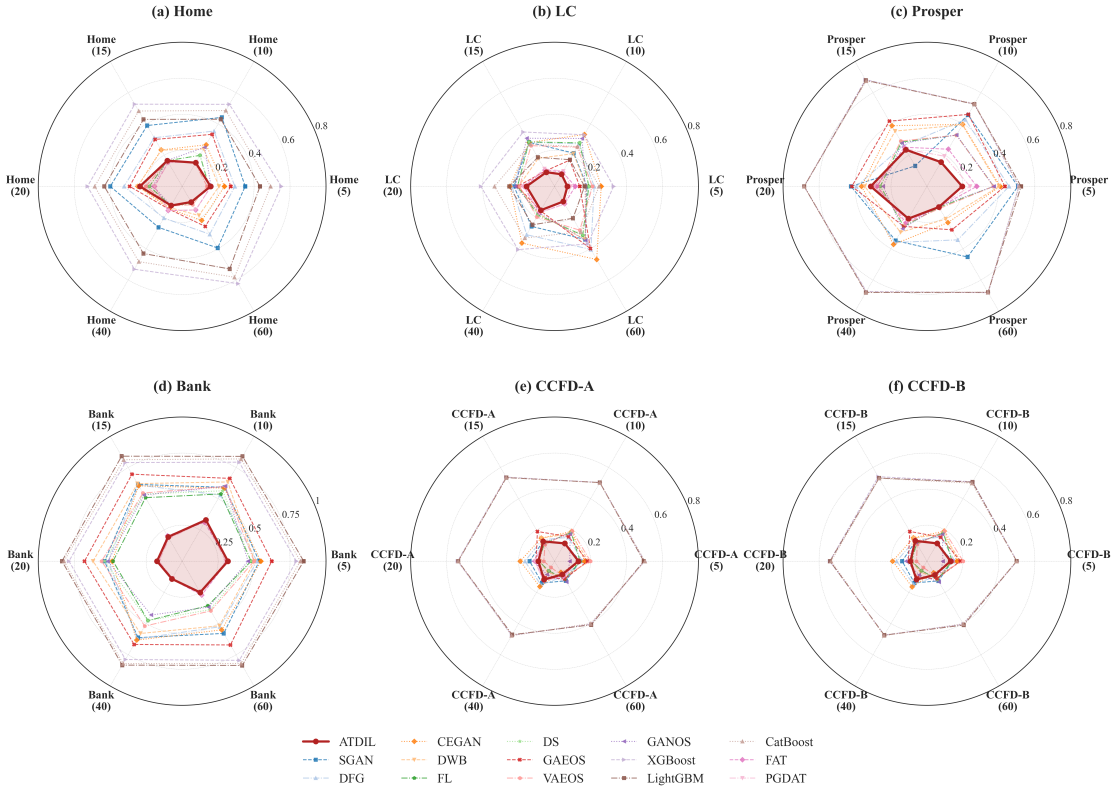

Figure A4 **Radar Visualization of SDS Values for Various Detection Models Across Six Datasets**

Table A2 Results of Ablation Experiments on Metrics AUROC and AUPRC

Dataset / Method	AUROC (%) (IR)						AUPRC (%) (IR)					
	5	10	15	20	40	60	5	10	15	20	40	60
Home												
ATDIL	**72.44**	**75.43**	**73.92**	**76.04**	**74.70**	**72.09**	**39.05**	**25.79**	**18.84**	**18.95**	**7.65**	**4.01**
CM-I	71.27	75.29	73.43	73.27	73.68	69.55	36.12	23.99	18.02	16.14	5.59	2.54
CM-II	71.97	72.38	72.41	72.55	71.61	70.92	37.93	25.11	18.55	17.30	6.02	2.30
LC												
ATDIL	**95.26**	**94.82**	**95.30**	**96.49**	**95.68**	**94.12**	**80.45**	**68.26**	**62.43**	**60.97**	**48.41**	**38.88**
CM-I	93.05	92.81	92.25	95.05	93.62	93.29	77.70	67.61	61.19	58.27	47.75	37.48
CM-II	94.46	94.65	95.18	94.79	93.21	92.87	78.87	67.94	60.28	58.80	45.44	38.41
CIS												
ATDIL	**90.03**	**89.58**	**90.39**	**90.01**	**89.73**	**88.56**	**74.83**	**65.86**	**61.94**	**57.52**	**49.79**	**40.23**
CM-I	87.11	86.78	87.26	88.62	87.68	88.35	72.04	64.90	61.62	55.29	47.24	37.26
CM-II	87.68	87.05	88.52	88.56	87.32	87.22	72.33	64.70	59.68	55.22	48.83	39.77
Prosper												
ATDIL	**74.61**	**72.32**	**72.62**	**73.69**	**71.78**	**72.89**	**37.20**	**21.37**	**14.94**	**13.18**	**6.67**	**4.45**
CM-I	71.95	70.35	71.70	70.44	69.98	70.03	35.92	20.03	13.44	11.30	4.55	2.30
CM-II	72.28	68.96	70.31	71.48	70.15	70.45	34.39	19.71	13.61	12.73	5.38	4.13
Bank												
ATDIL	**96.82**	**94.67**	**96.28**	**96.83**	**94.43**	**94.94**	**89.67**	**80.18**	**83.28**	**75.34**	**61.57**	**65.57**
CM-I	93.37	94.50	94.09	96.72	93.14	92.39	89.30	78.09	83.28	72.73	60.36	63.19
CM-II	94.11	91.63	95.37	95.95	92.11	91.83	87.69	79.46	80.64	74.07	60.72	65.29
CCFD-A												
ATDIL	**98.18**	**99.11**	**98.41**	**97.40**	**98.31**	**98.04**	**95.69**	**94.37**	**93.71**	**92.87**	**91.95**	**89.98**
CM-I	97.88	97.84	97.57	94.03	97.52	96.71	94.22	93.16	92.41	90.73	90.57	87.19
CM-II	94.88	96.04	96.90	96.58	98.25	96.79	94.73	92.09	93.04	90.06	91.64	88.37
CCFD-B												
ATDIL	**97.18**	**98.58**	**98.63**	**98.76**	**96.52**	**94.76**	**79.08**	**84.52**	**80.22**	**88.01**	**70.74**	**75.20**
CM-I	94.48	95.59	96.38	98.14	94.21	92.69	78.85	83.81	78.17	86.19	69.58	74.69
CM-II	96.65	95.64	95.14	95.66	95.24	94.33	77.83	81.76	79.16	85.91	70.25	73.76